EEL6507: Queueing Theory and Data Communications

Yuguang Fang

Department of Electrical and Computer Engineering
University of Florida
Tel: (352) 846-3043, Fax: (352) 392-0044
Email: *fang@ece.ufl.edu.edu*
Webpage: *http://www.fang.ece.ufl.edu*

1 Introduction to Communications Networks

1.1 Why do we need this course

This course is to present the methods for the design of computer communication networks to support many interesting applications. The tools developed in this course can be used for network design, service provisioning, and resource dimensioning.

Example 1: You just join a company, which is in the process of upgrading their systems. You are asked to evaluate the computer networks, create a budget plan for the company. What should you do?

Solution: Evaluate the traffic and potential future traffic, decide how much network capacity is necessary, then carry out a cost analysis, etc

Example 2: A major WAN service provider (e.g., ATT/Verizon) intends to install an optical fiber link between Gainesville and Orlando, you are responsible for the link design? What should you do?

Possible solution: Predict/analyze the traffic the link has to handle and the service characteristics, possibly the QoS requirements, then decide how many wavelengths you need in the waveguide.

Example 3: A wireless services provider just won an FCC auction bid for a chunk of frequency to be used, you are hired with big money to figure out how to build their wireless networks in Gainesville area. In one week, you have to propose the building plan, what should you do?

Possible solution: Figure out what you need: man power, resources, field study etc. You have to hire a traffic engineer to perform the traffic analysis!

Example 4: You study computer sciences and want to do a business on web hosting. How much money do you want to spend for your server?

1.2 Historical Overview

- Primitive forms of communications (e.g., smoke signals)
- Marconi's experiments on radio communications (wireless telegraphy) in 1897
- Armstrong's FM technology revolution in 1938

- Shannon's paper and information theory in 1948
- von Neuman's paper and invention of computers in 1948
- Central computers to remote terminals and other peripheral devices (1950s)
- Network with one central processors but with shared links
- Local area networks (LANs)
- Wide area networks (WANs)
- Internet
- Globalization of service integration
  - Wireless communications
  - Optical Communications

1.3 Communication Technology

- Wired: cable or optical fiber or powerline or phoneline — high speed
- Wireless: radio, microwave, infrared, satellite — convenience

1.4 Applications

- Remote access and computing, file transfer
- Remote update and transactions
- Electronic mail
- Telephony: voice services
- Video conferencing
- Internet surfing
- E-commerce/m-commerce
- Telemedicine

- Faxing via Internet, Internet Telephony (IP telephony)
- Internet gaming (Internet entertainment)
- Distance learning
- Digital library and digital government
- Web publishing
- Fight crimes: Woodstock'99 criminal identification
- More...*we could not live without...*

1.5 Information Transfer Units

- Message: independent data unit which has its meaning itself.
- Packet: Parts of a message, used for easier information transfer, the concept which revolutionized the data communications.

1.6 Sessions

- A process for fulfilling the communications between two end points or one point with many points.
- Modeling such processes are very important:
  - *Message arrivals*: the rate and the variability
  - *Session holding time*
  - *Message length*: mean and its distribution
  - *Allowable delay*: QoS
  - *Reliability*: error characteristics
  - *Message and packet ordering*: must be delivered in order

1.7 Switching

- Circuit switching (message switching)
- Packet switching (store-and-forward)– second revolution
- Virtual circuit switching: resource sharing idea—third revolution (ATM)

1.8 Layering

Layering, or the layered architecture, is a form of hierarchical modularity that is central to data network design—peering process design.

- *Physical layer*: raw bits transfer from point to point
- *Data link control layer*: reliable transfer of frames from point to point
  - Logical link layer: hiding physical media from high layer
  - Multiple access control (MAC): multiple users share a single link
- *Network layer*: choosing the right path to move packets around effectively and efficiently –routing and congestion control
- *Transport layer*: Packaging —dissembling/assembling, flow control, call admission control, reliability check in the message level
- *Session layer*: handling the interactions between two end points in setting up a session
- *Presentation layer*: data encryption, data compression, and code conversion
- *Application layer*: Yes, various applications–telnet, ftp

1.9 Network design

- Network throughput—network capacity
- Network delay — point-to-point, end-to-end
- Network dimensioning — resource (storage and transmission)
- QoS — voice, data, multimedia —more important than ever
- Connection blocking and dropping

2 Probability Refresher

2.1 The intuition of probability

It is the frequency of the same something happening, you could understand it as the "possibility", although it is NOT accurate mathematically.

Three components:

- A set of possible experimental outcomes;
- A group of these outcomes into classes called results
- The relative frequency of these results

2.2 Mathematical model

- A sample space S
- Events: a set of sets in S—$\mathcal{E}$
- Probability measure P: a nonnegative function defined on $\mathcal{E}$ satisfying: $0 \leq P(A) \leq 1$, $P(S) = 1$, $P(A \cup B) \leq P(A) + P(B)$ for any $A \in \mathcal{E}$ and $B \in \mathcal{E}$.

Remark: The union $\cup$ and the intersection $\cap$ are defined as usual.

Mutually exclusive: disjoint, $A \cap B = \emptyset$.

Exhaustive: $\bigcup_i^N A_i = S$.

Mutually exclusive and exhaustive: $\bigcup_i^N A_i = S$ and $A_i \cap A_j = \emptyset$.

Conditional Probability: absolute probability is the probability that no prior knowledge is known, while the conditional probability is the probability when some information is known.

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

2.3 Independence

A and B are called independent iff $P(A \cap B) = P(A)P(B)$, i.e., $P(A|B) = P(A)$.

2.4 Theorem of total probability

If $A_1, A_2, \ldots, A_N$ are mutually exclusive and exhaustive, then for any $B \in \mathcal{E}$, we have

$$P(B) = \sum_{i=1}^{N} P(A_i \cap B) = \sum_{i=1}^{N} P(B|A_i)p(A_i).$$

2.5 Bayes' Theorem

If $A_1, A_2, \ldots, A_N$ are mutually exclusive and exhaustive, then for any $B \in \mathcal{E}$, we have

$$P(A_i|B) = \frac{P(B|A_i)P(A_i)}{\sum_{j=1}^{N} P(B|A_j)P(A_j)}.$$

2.6 Random variables

Definition: $X(\omega)$ or simply X—a mapping from S to R (the set of real numbers) satisfying $(X \leq x) \in \mathcal{E}$.

Characterization: probability distribution, $(X = x) = \{\omega : X(\omega) = x\}$. In discrete case, $P(X = x_i) = p_i$. In continuous case, $P(X \leq x) = P(\omega : X(\omega) \leq x)$.

CDF: Cumulative distribution function— $(p_1, p_2, \ldots, p_K)$ for discrete case, $F(x) = P(X \leq x)$ for continuous case.

Properties of CDF: $0 \leq F(x) \leq 1$, $F(\infty) = 1$, $F(-\infty) = 0$, and $P(a < X \leq b) = F(b) - F(a)$, $F(b) \geq F(a)$ if $b \geq a$.

pdf: Probability density function (mass function)—

$$f(x) = \frac{dF(x)}{dx}$$

or

$$F(x) = \int_{-\infty}^{x} f(t)dt.$$

Property of pdf: $f(x) \geq 0$ and $\int_{-\infty}^{\infty} f(x)dx = 1$ and

$$P(a < X \leq b) = \int_{a}^{b} f(x)dx.$$

Example: exponential distribution

$$\begin{aligned} f(x) &= \lambda e^{-\lambda x},\ x \geq 0; \\ F(x) &= 1 - e^{-\lambda x},\ x \geq 0. \end{aligned}$$

2.7 Random vector and joint probability distribution

Random vector: $X = (X_1, X_2, \ldots, X_n)$ where X_i is a random variable, must satisfy $(X \in B) \in \mathcal{E}$ where B is a measurable set in $n-$dimensional Euclidean space R^n.

Joint distribution: For $x = (x_1, x_2, \ldots, x_n)$,

$$\begin{aligned} F(x) &= F(x_1, x_2, \ldots, x_n) = P(X_1 \leq x_1, X_2 \leq x_2, \ldots, X_n \leq x_n) \\ &= P[(X_1 \leq x_1) \cap (X_2 \leq x_2) \cap \cdots \cap (X_n \leq x_n)]. \end{aligned}$$

pdf:

$$f(x) = f(x_1, x_2, \ldots, x_2) = \frac{\partial^n F(x_1, x_2, \ldots, x_n)}{\partial x_1 \partial x_2 \cdots \partial x_n},$$

or

$$F(x_1, x_2, \ldots, x_n) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \cdots \int_{-\infty}^{x_n} f(t_1, t_2, \ldots, t_n) dt_n dt_2 \cdots dt_1.$$

Independence: $X_1, X_2, \ldots, X_n$ are independent iff

$$f(x_1, x_2, \ldots, x_n) = f_1(x_1) f_2(x_2) \cdots f_n(x_n).$$

Conditional probability density function:

$$f_{X|Y}(x|y) = \frac{d}{dx} P[X \leq x | Y = y] = \frac{f_{XY}(x, y)}{f_Y(y)}.$$

Function of a random variable: $Y = g(X)$,

$$F_Y(y) = P(Y \leq y) = P(\{\omega : g(X(\omega)) \leq y\}).$$

2.8 Expectation

Definition

$$E[\xi] = \sum_{i=1}^{n} \xi_i P(\xi = \xi_i) \quad (discrete)$$

$$\begin{aligned} E[\xi] &= \int_{-\infty}^{\infty} xf(x)dx = \int_{-\infty}^{\infty} xdF(x) \quad (continuous) \\ E[\phi(\xi)] &= \int_{-\infty}^{\infty} x\phi(x)f(x)dx = \int_{-\infty}^{\infty} \phi(x)dF(x) \quad (general) \\ E[\phi(X_1, X_2, \ldots, X_k)] &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} \phi(x_1, x_2, \ldots, x_k)dF(x_1, x_2, \ldots, x_k) \end{aligned}$$

Properties: linear, independence.

Concepts:

$$\begin{aligned} E[X^n] &= \int_{-\infty}^{\infty} x^n f(x)dx \quad \text{(nth moment)} \\ E[X - E(X)]^n &= \int_{-\infty}^{\infty} (x - E(X))^n f(x)dx \quad \text{(nth central moment)} \\ \sigma_X^2 &= E(X - E(X))^2 = E(X^2) - [E(X)]^2 \quad \text{(variance)} \\ C_X &= \frac{\sigma_X}{E(X)} \quad \text{(coefficent of variation)} \end{aligned}$$

2.9 Transforms

Characteristic function

$$\phi_X(u) = E[e^{juX}] = \int_{-\infty}^{\infty} e^{jux} f_X(x)dx.$$

Moment generating function:

$$M_X(v) = E[e^{vX}] = \int_{-\infty}^{\infty} e^{vx} f_X(x)dx.$$

Laplace transform of the pdf:

$$f^*(s) = E[e^{-sX}] = \int_{-\infty}^{\infty} e^{-sx} f_X(x)dx.$$

probability generating function: for discrete case

$$G(z) = E[z^X] = \sum_k z^k P(X = k).$$

Properties

1. $E[X^n] = (-j)^n \phi_X^{(n)}(0) = M_X^{(n)}(0) = (-1)^n f^{*(n)}(0)$.
2. $\Pr(X = k) = G^{(k)}(0)/k!$.
3. Let $\phi(u_1, u_2, \ldots, u_k) = E[e^{j(u_1X_1 + u_2X_2 + \cdots + u_kX_k)}]$ be the characteristic function of the random vector $(X_1, X_2, \ldots, X_k)$. Random variables $X_1, X_2, \ldots, X_k$ are independent iff $\phi(u_1, u_2, \ldots, u_k) = \phi_{X_1}(u_1)\phi_{X_2}(u_2)\cdots\phi_{X_k}(u_k)$. This statement is still valid when we use the moment generating function and the Laplace transform instead of characteristic function.

3 Markov Chain Theory

"Life is a Markov chain, the future depends on the current, but independent of the past." (MIT BBS)

3.1 Stochastic Processes

Definition: $X(t, \omega)$ or simply $X(t)$ is called stochastic process (random process) if it is a measurable mapping from the probability space $(S, \mathcal{E}, P)$ to the real line R: for each fixed t, $X(t)$ is a random variable and for each fixed ω, $X(t)$ is a measurable function on R.

Examples: stock price, the number of users in a computer system, the number of customers in a supermarket, the number of busy channels in a cell in a cellular network, etc.

Characterization: Distribution $F_X(x,t) = P(X(t) \leq x)$; Joint distribution function: for any $t_1, t_2, \ldots, t_n$ and $x_1, x_2, \ldots, x_n$,

$$F_X(\mathbf{x}, \mathbf{t}) = F_{X_1 X_2 \cdots X_n}(x_1, x_2, \ldots, x_n; t_1, t_2, \ldots, t_n) = P(X(t_1) \leq x_1, X(t_2) \leq x_2, \ldots, X(t_n) \leq x_n).$$

where $\mathbf{x} = (x_1, x_2, \ldots, x_n)$ and $\mathbf{t} = (t_1, t_2, \ldots, t_n)$.

Stationarity: A process $X(t)$ is said to be *stationary* if for any τ,$t_1, t_2, \ldots, t_n$ and $x_1, x_2, \ldots, x_n$, we have

$$P(X(t_1+\tau) \leq x_1, X(t_2+\tau) \leq x_2, \ldots, X(t_n+\tau) \leq x_n) = P(X(t_1) \leq x_1, X(t_2) \leq x_2, \ldots, X(t_n) \leq x_n),$$

i.e., the joint distribution function is time-shift invariant.

Correlation: mean $E(X(t))$, autocorrelation:

$$R_{XX}(t_1, t_2) = E[X(t_1)X(t_2)].$$

Wide-sense stationarity: A random process $X(t)$ is *wide-sense stationary* if its mean and autocorrelation functions are time-shift invariant, $E[X(t)] = c$ and $R_{XX}(t_1, t_2) = R_{XX}(t_2 - t_1)$.

3.2 Markov Processes

General Interpretation: A random process is called a *Markov process* if the future does not depend on the past when the current information is available, i.e., when the current information is known, the future and the past are independent.

Intuitive examples: The number of customers in the supermarket; the number of call arrivals in a cell in cellular systems. The non-Markov process: the long term investment returns, the web traffic.

Classification: The set of value a process can assume is called the *state space*. Both time and space can be either discrete and continuous, this leads to *discrete-time Markov process* and *continuous-time Markov process*. If the state space is a discrete set (either finite or infinite), the Markov process is called *Markov chain*. In this course, we concentrate on the Markov chain only.

3.3 Discrete-time Markov process

Definition: $\{X_n\}$ is a discrete time random process, if for any integer m and any measurable sets $A_{n+1}, A_n, \ldots, A_{n-m}$, we have the following

$$P(X_{n+1} \in A_{n+1} | X_n \in A_n, \ldots, X_{n-m} \in A_{n-m}) = P(X_{n+1} \in A_{n+1} | X_n \in A_n),$$

then we call this process the discrete-time Markov process.

Discrete-time Markov chain: A random process X_n with discrete state space is called *discrete-time Markov chain* if for any integers m and any states $x_{n+1}, x_n, \ldots, x_{n-m}$, we have

$$P(X_{n+1} = x_{n+1} | X_n = x_n, X_{n-1} = x_{n-1}, \ldots, X_{n-m} = x_{n-m}) = P(X_{n+1} = x_{n+1} | X_n = x_n).$$

If the state space is finite set, the the process is called *finite state Markov chain*.

Characterization of finite state Markov chain: the probability transition matrix: $P = (p_{ij}(n))$ where

$$p_{ij} = P(X_{n+1} = j | X_n = i).$$

If $p_{ij}(n)$ does not depend on the time n, i.e., $p_{ij}(n) = p_{ij}$, then the chain is called *homogeneous finite state Markov chain*. We will concentrate on homogeneous Markov chain only. The $m-$step transition probability is defined as

$$p_{ij}^{(m)} = P(X_{n+m} = j | X_n = i).$$

From the Theorem of total probability, we have the following Chapman-Kolmogrov equations:

$$p_{ij}^{(m)} = \sum_k p_{ik}^{(m-1)} p_{kj} = \sum_k p_{ik} p_{kj}^{(m-1)}.$$

More generally, for any $0 < l \leq m$

$$p_{ij}^{(m)} = \sum_k p_{ik}^{(l)} p_{kj}^{(m-l)},$$

which is equivalent to

$$P^m = P^l \cdot P^{m-l}.$$

A state j is said to be *reachable* from state i if there exists an integer m_0 such that $p_{ij}^{(m_0)} > 0$. A Markov chain is said to be *irreducible* if every state is reachable from any other state. We say that a state i is *periodic* if there exists some integer $m \geq 1$ such that $p_{ii}^{(m)} > 0$ and some integer $d > 1$ such that $p_{ii}^{(n)} > 0$ only if n is the multiple of d. A Markov chain is said to be *aperiodic* if none of its states is periodic. A probability distribution $\{p_j | j \geq 0\}$ is said to be *stationary distribution* for the Markov chain if

$$p_j = \sum_i p_i p_{ij}, j \geq 0,$$

or equivalently,

$$\pi = \pi P, \;\; \pi = (p_0, p_1, p_2, \ldots).$$

Let

$$p_j = \lim_{n\to\infty} P(X_n = j | X_0 = i), \; i \geq 0$$

(which exists and is independent of initial state i for irreducible and aperiodic Markov chain). It can also be shown that

$$p_j = \lim_{k\to\infty} \frac{\text{number of visits to state j up to time k}}{k}$$

which implies that p_j is the proportion of time or the frequency with which the process visits j, a time average interpretation.

Fundamental Theorem of Markov chain: In an irreducible, aperiodic Markov chain, there are two possibilities for the scalars $p_j = \lim_{n\to\infty} P(X_n = j | X_0 = i)$:

(1) $p_j = 0$ for all $j \geq 0$, in which case the chain does not have stationary distribution;

(2) $p_j > 0$ for all $j \geq 0$, in which case $\{p_j | j \geq 0\}$ is the unique stationary distribution of the chain, i.e., it satisfies the following set of equations: $\pi = \pi P$ and $\pi \cdot e = 1$ where e is a column vector with all entries equal to unity.

Remarks: Case (1): the number of customers in a queueing system where the arrival rate is greater than the service rate. Case (1) never arises for finite state Markov chain!

Global balance equation:

$$\sum_{i=0}^{\infty} p_j p_{ji} = \sum_{i=0}^{\infty} p_i p_{ij},$$

which implies that the flow-out probability (left hand side) is equal to flow-in probability (right hand side). A generalized version is as follows:

$$\sum_{j\in I}\sum_{i\notin I} p_j p_{ji} = \sum_{i\notin I}\sum_{j\in I} p_i p_{ij}$$

where I is a subset of states.

Detailed balance equations: Sometimes, the probability transitions only occurs to the neighbors such as in birth-death process, in which case we have

$$p_i p_{ij} = p_j p_{ji},\ i,j \geq 0$$

this is called *detailed balance equation*.

Partial balance equations: For every state j, consider a partition $S_j^1, \ldots, S_j^k$ of the complementary set of states $\{i | i \geq 0, i \neq j\}$, the partial balance equations are given

$$p_j \sum_{i\in S_j^m} p_{ji} = \sum_{i\in S_j^m} p_i p_{ij},\ m = 1, 2, \ldots, k.$$

3.4 Continuous-time Markov chain

Definition: A random process $X(t)$ with discrete state space is called a *continuous-time Markov chain* if for any time instants $t_1 \leq t_2 \leq \cdots \leq t_n \leq t_{n+1}$ and states $i_1, i_2, \ldots, i_n, i_{n+1}$, we have

$$P[X(t_{n+1}) = i_{n+1} | X(t_n) = i_n, \ldots, X(t_1) = i_1] = P[X(t_{n+1}) = i_{n+1} | X(t_n) = i_n].$$

An equivalent characterization is the following:

(a) *Exponential sojourn time*: the time the process spends in any state i is exponentially distributed with parameter v_i;

(b) *Markov jump process*: when the process leaves the state i, it will enter state j with probability p_{ij}, where $\sum_j p_{ij} = 1$.

Define

$$q_{ij} = \nu_i p_{ij},\ \ i,j \geq 0$$

which can be shown to be the transition rate from state i to state j. Assume that the *embedded Markov chain* is irreducible. Let

$$p_{ij}(s,t) = P(X(t) = j | X(s) = i),\ H(s,t) = (p_{ij}(s,t)),$$

then we have the following Chapman-Kolmogrov equations

$$p_{ij}(s,t) = P(X(t) = j | X(s) = i) = \sum_k p_{ik}(s,u) p_{kj}(u,t)$$

i.e.,

$$H(s,t) = H(s,u)H(u,t),\ H(t,t) = I$$

for any $s \leq u \leq t$ and I is the identity matrix. If the probability transition matrix $H(s,t)$ does not depend on the time, i.e., $H(s,t)$ is a function of the time difference $t-s$, then the chain is called *homogeneous* Markov chain. As in the discrete-time case, we have

$$p_{ij}(s, t+\Delta t) = P(X(t+\Delta t) = j | X(s) = i) = \sum_k p_{ik}(s,t) p_{kj}(t, t+\Delta t)$$

or

$$H(s, t+\Delta t) = H(s,t)H(t, t+\Delta t),$$

thus we have

$$H(s, t+\Delta t) - H(s,t) = H(s,t)[H(t, t+\Delta t) - H(t,t)] = H(s,t)[H(t, t+\Delta t) - I],$$

by dividing both sides by Δt and by letting $\Delta t \to 0$, we obtain the *forward Chapman-Kolmogorov equation*

$$\frac{\partial H(s,t)}{\partial t} = H(s,t)Q(t)$$

where

$$Q(t) = \lim_{\Delta \to 0} \frac{H(t, t+\Delta t) - I}{\Delta},$$

i.e.,

$$\begin{aligned} q_{ii}(t) &= \lim_{\Delta \to 0} \frac{p_{ii}(t, t+\Delta t) - 1}{\Delta t} \\ q_{ij}(t) &= \lim_{\Delta \to 0} \frac{p_{ij}(t, t+\Delta t)}{\Delta t},\ i \neq j \end{aligned}$$

Remark: We have $p_{ii}(t, t+\Delta) - 1 = q_{ii}(t)\Delta + o(\Delta t)$, i.e., $1 - p_{ii}(t, t+\Delta t) = -q_{ii}(t)\Delta t + o(\Delta t)$, which implies that the departure rate from state i is $-q_{ii}(t)$, i.e., the "service time" or the sojourn time at state i is exponentially distributed. We also have $p_{ij}(t, t+\Delta t) = q_{ij}(t)\Delta t + o(\Delta t)$ $(i \neq j)$.

Flow rate equations:

$$\sum_j q_{ij} = 0,\ \ i \geq 0.$$

Similarly, we have the *backward Chapman-Kolmogrov equation*

$$\frac{\partial H(s,t)}{\partial s} = -Q(s)H(s,t),\ \ s \leq t.$$

We also have

$$H(s,t) = e^{\int_s^t Q(u)du}.$$

Probability distribution: Let

$$p_j(t) = P(X(t) = j), \;\; j \geq 0 \text{ and} \pi(t) = (p_0(t), p_1(t), \ldots).$$

Given any initial distribution $\pi(0)$, we have

$$\pi(t) = \pi(0)H(0,t) = \pi(0)e^{\int_0^t Q(u)du},$$

or

$$\frac{d\pi(t)}{dt} = \pi(t)Q(t).$$

Homogeneous Markov chain: If the Markov chain is homogeneous, i.e., $H(s,t)$ only depends on the time difference $t-s$, we can have simpler results. Let

$$\begin{aligned} p_{ij}(t) &= p_{ij}(s, s+t) \\ q_{ij} &= q_{ij}(t), \; i,j \geq 0 \\ H(t) &= H(s, s+t) = (p_{ij}(s, s+t)) \\ Q &= Q(t) = (q_{ij}) \end{aligned}$$

We have

$$\begin{aligned} H(s+t) &= H(s)H(t) \\ \frac{dH(t)}{dt} &= H(t)Q = QH(t), \;\; H(0) = I \\ H(t) &= e^{Qt} \\ \frac{d\pi(t)}{dt} &= \pi(t)Q = \pi(0)e^{Qt} \end{aligned}$$

If the chain is irreducible, we have

$$p_j = \lim_{t\to\infty} P(X(t) = j | X(0) = i) = \lim_{t\to\infty} \frac{T_j(t)}{t},$$

where $T_j(t)$ is the time spent in state j up to time t.

Fundamental Theorem of Markov chain: For an irreducible homogeneous Markov chain, the limits

$$p_j = \lim_{t\to\infty} p_{ij}(t) = \lim_{t\to\infty} p_j(t)$$

always exist, is independent of the initial state i and satisfy the following equations:

$$\pi Q = 0, \;\; \pi \cdot e = 1,$$

where $\pi = (p_0, p_1, p_2, \ldots)$ and $e = (1, 1, 1, \ldots)^T$ (T denotes the transpose).

global balance equations:

$$p_j \sum_i q_{ji} = \sum_i p_i q_{ij}, \; j \geq 0$$

i.e., the flow rate out of state j is equal to flow rate into state j.

Detailed balance equations:

$$p_j q_{ji} = p_i q_{ij}, \;\; i, j \geq 0.$$

Homework # 1

1. The exponential distribution for the random variable ξ has the following density function

$$f(t) = \lambda e^{-\lambda t},\ t \geq 0.$$

   (1). Find the mean, the variance and the coefficient of variation;

   (2). Find the characteristic function, moment generating function and the Laplace transform of the pdf;

   (3). Show that the exponential distribution has the following memoryless property: for any $s \geq 0$ and $t \geq 0$, we have

$$P(\xi \geq s + t | \xi \geq s) = P(\xi \geq t).$$

   In fact, a stronger version can be shown: for any nonnegative random variable ν, we have

$$P(\xi \geq t + \nu | \xi \geq \nu) = P(\xi \geq t).$$

2. Find the pdf for the smallest of K independent random variables, each of which is exponentially distributed with parameter λ. Find the Laplace transform of the pdf for the sum of the K independent random variables, each of which is exponentially distributed?

3. Consider the discrete-time, discrete-state Markov chain with the probability transition matrix

$$\begin{pmatrix} 0.5 & 0.5 \\ 0.75 & 0.25 \end{pmatrix}.$$

   Find the stationary state probability vector π. Suppose that you are an outsider observer and observe the evolution of the Markov chain, how often you find the chain stay in the state 1?

4 Queueing Systems Basics and Little's Theorem

4.1 Introduction

Examples

- Supermarket checkout model: multiple server systems
- Factory assembling line: tandem queues
- Cellular networks: a single cell model
- Web server: time-sharing systems
- Cloud server model and interconnected cloud servers (cloud center)
- IP phone services
- Analysis of medium access control protocols and ARQ protocols
- . . .

Common features

- Arrival process (jobs)
- Serving process (servers)
- Service discipline (ordering)

Performance Indices

- Number of jobs/customers in the systems
- Delay per customer/overall delay (network delay)
- Throughput in shared network
- Blocking probability/loss probability when resource is limited

Why are we interested?: network design

- Network dimensioning
- QoS: delay requirement or loss probability

What do we need?

- The customer/job arrival process: interarrival time distribution
- The service time distribution
- Service disciplines: FIFO

How do we conduct the performance evaluation?

- Queueing theory (analytical approach)
- Simulations

4.2 Classification: Kendall's notation

Kendall notation

- Arrival process/service process/# of servers/# of buffers/population model
- G/G/m/m/F
  - G—general distribution
  - M—Markov process/exponential distribution
  - D—constant distribution
  - F—finite population
  - ∞ — when $m = \infty$, the notation will be omitted
- M/M/1: Poisson arrival, exponential service time, single server and infinite buffer
- M/G/1: Poisson arrival, general service time and single server
- G/M/1: general arrival, exponential service time and single server
- M/M/1/m: Poisson arrival, exponential service time, finite buffer

- M/M/m/m: Poisson arrival, exponential service time, finite servers and finite buffers
- G/G/1: general arrival, general service time and single server

Remark: In all notations, the arrival process and the service process are assumed to be independent.

4.3 Exponential distribution and Poisson processes

Exponential distribution: The most important distribution in computer networks.

Let X be distributed with the following exponential distribution:

$$F_X(x) = P(X \leq x) = 1 - e^{-\mu x}, \;\; f_X(x) = \mu e^{-\mu x}, \;\; x \geq 0.$$

Memoryless property: $\forall r, t \geq 0$, we have

$$P(X \geq r + t | X \geq t) = P(X \geq r).$$

Strong Memoryless Property (Fang 1999): $\forall$ random variable $\xi \geq 0$ and $\forall$ time $r \geq 0$, we have

$$P(X \geq \xi + r | X \geq \xi) = P(X \geq r).$$

Proof: Let $1/\mu = E[X]$ and let $f_\xi(y)$ denote the probability density function of ξ, then we have

$$\begin{aligned} \Pr(X \geq \xi + r | X \geq \xi) &= \frac{\int_0^\infty \Pr(X \geq y + r) f_\xi(y) dy}{\Pr(X \geq \xi)} \\ &= \frac{\int_0^\infty e^{-\mu(y+r)} f_\xi(y) dy}{\Pr(X \geq \xi)} = \frac{e^{-\mu r} \int_0^\infty e^{-\mu y} f_\xi(y) dy}{\Pr(X \geq \xi)} = e^{-\mu r}. \end{aligned}$$

Poisson distribution

Definition A random variable X is *Poisson distributed* with parameter a if X is taken integer values and has the following probability distribution:

$$P(X = n) = \frac{a^n}{n!} e^{-a}, \;\; n = 0, 1, 2, \ldots$$

Mean and Variance: Its mean and variance are given by

$$\overline{X} = a, \;\; \sigma_X^2 = a.$$

Moment generating function

$$p(z) = E(z^X) = e^{a(z-1)}.$$

4.4 Poisson Process

This process arose from the POTS (Plain Old Telephone Systems), which is used to describe the arrival process.

Intuitive definition: A Poisson arrival process is equivalent to the following descriptive characterization:

- For any sufficiently small time interval, the probability that there is one arrival is proportional to the length of the interval;
- The probability that there are more than one arrival in a sufficiently small interval can be negligible;
- The number of arrivals are incrementally independent, i.e., the number of arrivals in one interval is independent of the number of arrivals in other non-overlapping interval.

Formal Definition: A random process $\{A(t)|t \geq 0\}$ is said to be a *Poisson process* with parameter λ if it takes nonnegative integers and satisfies the following properties:

(a) $A(t)$ is a counting process that represents the total number of arrivals occurred from zero to t, and for any $t > s \geq 0$, $A(t) - A(s)$ equals the number of arrivals in the interval $(s, t]$;

(b) $A(t)$ is incrementally independent, i.e., $A(b) - A(a)$ is independent of $A(d) - A(c)$ whenever $a \leq b < c \leq d$;

(c) For any $t, \tau > 0$, we have

$$P(A(t+\tau) - A(t) = n) = \frac{(\lambda\tau)^n}{n!} e^{-\lambda\tau}, \;\; n = 0, 1, 2, \ldots$$

i.e., $A(t+\tau) - A(t)$ is Poisson distributed with parameter $\lambda\tau$.

Properties of the Poisson process A(t):

(1). Poisson process can be fully characterized by the interarrival time

$$\tau_n = t_n - t_{n-1}$$

where t_n denotes the $n-$th arrival time instant, i.e., $A(t)$ is Poissonian iff τ_n is exponentially distributed with the same parameter λ.

(2). $\forall t \geq 0,\ \delta \geq 0$, we have

$$\begin{aligned} P(A(t+\delta) - A(t) = 0) &= 1 - \lambda\delta + o(\delta) \\ P(A(t+\delta) - A(t) = 1) &= \lambda\delta + o(\delta) \\ P(A(t+\delta) - A(t) \geq 2) &= o(\delta) \end{aligned}$$

(3). If two or more independent Poisson processes $A_1(t), A_2(t), \ldots, A_k(t)$ are merged into a single process $A(t) = A_1(t) + A_2(t) + \cdots + A_k(t)$, then $A(t)$ is also a Poisson process with parameter equal to the sum of the parameters.

(4). Independent splitting also leads to Poisson processes, i.e., an arrival from a Poisson process $A(t)$ with parameter λ will belong to the i arrival stream with probability p_i where $\sum_{i=1}^{k} p_i = 1$, then each individual arrival stream is also a Poisson process with parameter $p_i\lambda$.

(5). A counting incrementally independent process $A(t)$ is a Poisson process iff its moment generation function is

$$p(z) = e^{\lambda t(z-1)}.$$

As a remark, we provide a less rigorous proof for the equivalence of the intuitive definition and the formal definition below.

Proof: Formal definition can be easily proved by linear approximation to the exponential function. Here we only need to show that the intuitive definition implies the formal definition. We only need to prove item (c), i.e., the Poisson property. Let slice time interval $[t, t+\tau]$ into n equal interval of length $\Delta = \tau/n$ and $t_i = t + i\Delta$. When n is very large, each shorter interval would contain either one arrival or none while the probability that there are more than arrivals can be negligible comparing to the probability with zero arrival or one arrival. Thus, we have (the equal size in the equation below indicates approximately equal when n is sufficiently small and $x_i = 0$ or $x_i = 1$)

$$\begin{aligned} &\Pr(A(t+\tau) - A(t) = k) \\ &= \sum_{x_1+\cdots+x_n=k} \Pr\left(A(t2) - A(t_1) = x_1, A(t_3) - t_2) = x_2, \cdots, A(t_n) - A(t_{n-1}) = x_n\right) \\ &= \sum_{x_1+\cdots+x_n=k} \prod_{i=1}^{n} \Pr\left(A(t_i) - A(t_{i-1}) = x_i\right) \\ &= \sum_{x_1+\cdots+x_n=k} \prod_{i=1}^{n} (\lambda\Delta)^{x_i}(1-\lambda\Delta)^(1-x_i) \\ &= \sum_{x_1+\cdots+x_n=k} (\lambda\Delta)^{x_1+\cdots+x_n}(1-\lambda\Delta)^{n-(x_1+\cdots+x_n)} \end{aligned}$$

$$= \left(\sum_{x_1+\cdots+x_n=k} 1 \right) (\lambda\Delta)^k (1-\lambda\Delta)^{n-k}$$
$$= \binom{n}{k} (\lambda\Delta)^k (1-\lambda\Delta)^{n-k}$$

which can be easily shown, when taking $\Delta = \tau/n$ into the equations, to converge to the Poisson distribution item (c) by letting $n \to \infty$. □

4.5 Little's Law

Little's law (or Little's Theorem) is a result expressing the average number of customers and the average time!

- $N(t)$ — # of customers in the system at time t
- $\alpha(t)$ — # of customers arriving in the interval $(0, t]$
- $\beta(t)$ — # of customers departing in the interval $(0, t]$
- T_i — time spent in the system by the $i-$th customer in the system

Interpretation of time average: for any time function $f(t)$, we have

$$\frac{1}{t}\int_0^\infty f(\tau)d\tau \approx \sum f(t_i)\frac{\Delta t_i}{t}$$

which is the time average!

Define

$$\begin{aligned}
N_t &= \frac{1}{t}\int_0^t N(\tau)d\tau \\
N &= \lim_{t\to\infty} N_t \text{ (average customer in the system)} \\
\lambda_t &= \frac{\alpha(t)}{t} \\
\lambda &= \lim_{t\to\infty} \lambda_t \text{ (average arrival rate)} \\
T_t &= \frac{\sum_{i=0}^{\alpha(t)} T_i}{\alpha(t)} \\
T &= \lim_{t\to\infty} T_t \text{ (average system time)}
\end{aligned}$$

Little's Law

$$N = \lambda T.$$

Intuition:

$$N = \frac{T}{1/\lambda} = \frac{\text{average system time per customer}}{\text{average interarrival time}}.$$

Proof: Let $\alpha(t)$ and $\beta(t)$ denote the arrivals and departures in the time interval $(0, t]$, respectively,

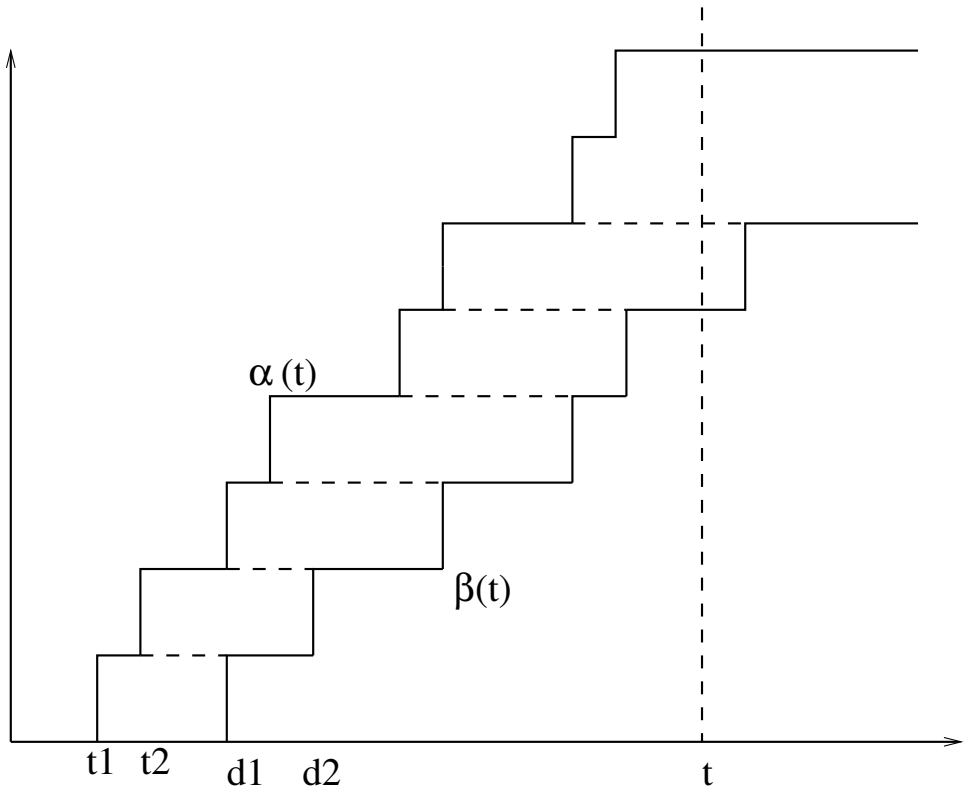


Figure 1: Little's law

then $N(t) = \alpha(t) - \beta(t)$. Thus, from Figure 1, we have

$$\begin{aligned}
\int_0^t N(\tau)d\tau &= \int_0^t [\alpha(\tau) - \beta(\tau)]d\tau \\
&= \int_0^t \alpha(\tau)d\tau - \int_0^t \beta(\tau)d\tau \\
&= \sum_{i=0}^{\alpha(t)} T_i
\end{aligned}$$

hence,

$$\frac{1}{t}\int_0^t N(\tau)d\tau = \frac{1}{t}\sum_{i=0}^{\alpha(t)} T_i = \frac{\alpha(t)}{t} \cdot \frac{\sum_{i=0}^{\alpha(t)} T_i}{\alpha(t)}.$$

Taking limit $t \to \infty$ on both sides, we have

$$N = \lambda T.$$

Probabilistic form of Little's Law

From the Law of Large Numbers (assuming all processes involved are Ergodic), or the following principle from physics:

$$time\ average\ =\ ensemble\ average,$$

we obtain

$$\lim_{t\to\infty}\frac{1}{t}\int_0^t N(\tau)d\tau) = E[N(t)] = \overline{N(t)} = \overline{N}$$
$$\lim_{t\to\infty}\frac{\alpha(t)}{t} = \left[\lim_{t\to\infty}\frac{\sum_{i=0}^{\alpha(t)}\tau_i}{\alpha(t)}\right]^{-1} = \frac{1}{E[\tau_i]} = \frac{1}{\overline{\tau}}$$
$$\lim_{t\to\infty}\frac{\sum_{i=0}^{\alpha(t)}T_i}{\alpha(t)} = E[T_i] = \overline{T}$$

hence, we have

$$\overline{N} = \frac{\overline{T}}{\overline{\tau}} = \lambda\overline{T},\ \lambda = 1/\overline{\tau}.$$

Example: Packet arrivals with rate λ, packet transmission time is $\overline{X}$. Let W denote the waiting time (not including the transmission time), the Little's Law gives

$$N_Q = \lambda W,$$

where N_Q denotes the average number of packets in the queue. The Little's law states that the relationship between the length of a time interval and the customers arriving in that time interval. Similarly, the average number of packets under transmission will be

$$\rho = \lambda\overline{X},$$

which is also called the *traffic intensity*.

4.6 Birth-Death Processes

A B-D process is a special case of Markov chain, in which transitions only occur to the neighbors.

- Birth $\leftrightarrow$ population increases only by "1"
- Death $\leftrightarrow$ population decreases only by "1"

We only consider the continuous-time B-D processes with discrete state space. Examples: the number of users in a computer network. The B-D processes are the most commonly used processes in queueing systems.

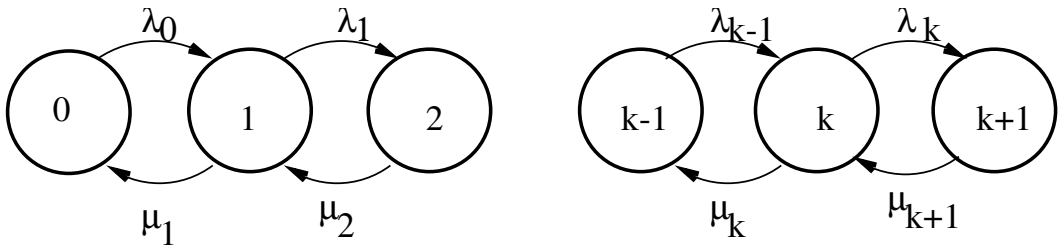
Definition: A B-D process is a Markov chain in which the state transitions from any state are permitted ONLY to its neighbors.

State space: $S = \{0, 1, 2, 3, \ldots\}$ — finite or infinite.

Transition rate matrix:

$$Q = \begin{pmatrix} -\lambda_0 & \lambda_0 & 0 & 0 & \ldots & \\ \mu_1 & -(\lambda_1+\mu_1) & \lambda_1 & 0 & 0 & \ldots \\ 0 & \mu_2 & -(\lambda_2+\mu_2) & \lambda_2 & 0 & \ldots \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ldots \end{pmatrix}.$$

State transition diagram:



Interpretation–birth-death

- λ_k—the birth rate when the population is k
- μ_k—the death rate when the population is k

Interpretation–flow

- λ_k—the flow rate from state k to state $k+1$
- μ_k—the flow rate from state k to state $k-1$

Interpretation–probabilistic: Let $N(t)$ denote the population at time t.

- B_1: (exactly 1 birth in $(t, t+\Delta t)|N(t) = k$), $P(B_1) = \lambda_k \Delta t + o(\Delta t)$

- D_1: (exactly 1 death in $(t, t+\Delta t)|N(t)=k$), $P(D_1) = \mu_k \Delta t + o(\Delta t)$
- B_2: (no birth in $(t, t+\Delta t)|N(t)=k$), $P(B_2) = 1 - \lambda_k \Delta t + o(\Delta t)$
- D_2: (no death in $(t, t+\Delta t)|N(t)=k$), $P(D_2) = 1 - \mu_k \Delta t + o(\Delta t)$
- Remarks: $P(\text{multiple births}) = P(\text{multiple deaths}) = o(\Delta t)$

Probability distribution of population

Let $N(t)$ be the B-D process, need to find

$$p_k(t) = P(N(t) = k) = \pi_k(t),$$

i.e., the probability that the population at time t is k.

First Method

Recall $\frac{d\pi(t)}{dt} = \pi(t)Q$, then $\pi(t) = \pi(0)e^{Qt}$. From this, we can obtain

$$\begin{aligned} \frac{dp_k(t)}{dt} &= -(\lambda_k + \mu_k)p_k(t) + \lambda_{k-1}p_{k-1}(t) + \mu_{k+1}p_{k+1}(t) \\ \frac{dp_0(t)}{dt} &= -\lambda_0 p_0(t) + \mu_1 p_1(t) \end{aligned}$$

Second Method

In order to find $P(N(t) = k)$, we need to find what leads to $N(t) = k$. The population is in state k at time interval $(t, t+\Delta t)$ if

(a) the population at time t is k and no birth in $(t, t+\Delta t)$

(b) the population at time t is $k-1$ and one birth in $(t, t+\Delta t)$

(c) the population at time t is $k+1$ and one death in $(t, t+\Delta t)$

and transitions from all other events (from multiple births or from multiple deaths) are negligible (i.e., $o(\Delta t)$). Thus, we have

$$\begin{aligned} p_k(t+\Delta t) &= P(N(t+\Delta t) = k) \\ &= P(N(t+\Delta t) = k|N(t) = k-1)P(N(t) = k-1) \\ &\quad P(N(t+\Delta t) = k|N(t) = k)P(N(t) = k) \\ &\quad P(N(t+\Delta t) = k|N(t) = k+1)P(N(t) = k+1) \\ &= p_{k-1,k}(\Delta t)p_{k-1}(t) + p_{k,k}(\Delta t)p_k(t) + p_{k+1,k}(\Delta t)p_{k+1}(t) \end{aligned}$$

When $k > 0$, we have

$$\begin{aligned} p_k(t+\Delta t) &= P(B_1)p_{k-1}(t) + P(B_2 \cap D_2)p_k(t) + P(D_1)p_{k+1}(t) \\ & [\lambda_{k-1}\Delta t + o(\Delta t)]p_{k-1}(t) + [1 - (\lambda_k + \mu_k)\Delta t + o(\Delta t)]p_k(t) + [\mu_{k+1}\Delta t + o(\Delta t)]p_{k+1}(t) \end{aligned}$$

hence

$$\frac{dp_k(t)}{dt} = -(\lambda_k + \mu_k)p_k(t) + \lambda_{k-1}p_{k-1}(t) + \mu_{k+1}p_{k+1}(t).$$

Similarly, we can obtain the case when $k = 0$. In conclusion, we have

$$\begin{aligned} \frac{dp_k(t)}{dt} &= -(\lambda_k + \mu_k)p_k(t) + \lambda_{k-1}p_{k-1}(t) + \mu_{k+1}p_{k+1}(t) \\ \frac{dp_0(t)}{dt} &= -\lambda_0 p_0(t) + \mu_1 p_1(t) \end{aligned}$$

Flow interpretation

Flow rate into state k=$\lambda_{k-1}p_{k-1}(t) + \mu_{k+1}p_{k+1}(t)$

Flow rate out of state k=$(\lambda_k + \mu_k)p_k(t)$

Hence, the effective probability transition rate is the difference between the flow-in rate minus the flow-out rate.

Solutions

It is difficult to solve these differential-difference equations! A standard procedure for tackling a difficult problem: attack simple cases first, start always with simple examples.

I). *Pure birth case:* $\mu_k = 0,\ \lambda_k = \lambda,\ \forall k$

In this case, the equations are as follows:

$$\begin{aligned} \frac{dp_k(t)}{dt} &= -\lambda p_k(t) + \lambda p_{k-1}(t), \quad k \geq 1 \\ \frac{dp_0(t)}{dt} &= -\lambda p_0(t) \end{aligned}$$

Initial condition: $p_k(0) = 1$ when $k = 0$ and $p_k(0) = 0$ when $k \neq 0$.

Solution: Solve the second equation for $p_0(t)$, which is given by $p_0(t) = e^{-\lambda t}$. Then recursively solve the equation, the final solution is

$$p_k(t) = \frac{(\lambda t)^k}{k!}e^{-\lambda t}.$$

Yes, it is a Poisson process! Intuition?

II). *Pure death process:* $\lambda_k = 0, \mu_k = \mu$, $0 < k \leq N$

$$\begin{aligned}
\frac{dp_k(t)}{dt} &= -\mu p_k(t) + \mu p_{k+1}(t), \ \ 0 < k < N \\
\frac{dp_N(t)}{dt} &= -\mu p_N(t) \\
\frac{dp_0(t)}{dt} &= \mu p_1(t)
\end{aligned}$$

The solution is given by

$$\begin{aligned}
p_k(t) &= \frac{(\mu t)^{N-k}}{(N-k)!} e^{-\mu t}, \ \ 0 < k \leq N \\
\frac{dp_0(t)}{dt} &= \frac{\mu(\mu t)^{N-1}}{(N-1)!} e^{-\mu t}
\end{aligned}$$

III). *B-D process:* $\lambda_k = \lambda$, $\mu_k = \mu$, $k > 0$, $\lambda_0 = \lambda$ $\mu_0 = 0$

$$\begin{aligned}
\frac{dp_k(t)}{dt} &= -(\lambda + \mu) p_k(t) + \lambda p_{k-1}(t) + \mu p_{k+1}(t), \ k > 0 \\
\frac{dp_0(t)}{dt} &= -\lambda p_0(t) + \mu p_1(t)
\end{aligned}$$

Facts:

- probability distribution $\leftrightarrow$ moment generating function
- Differential equations $\leftrightarrow$ Laplace transform operations

Idea: Find moment generating function and Laplace transform. From ODE: Laplace transform can be used to find solution of ODE (Ordinary Differential Equations).

Procedure: Let

$$p(z,t) = \sum_{k=0}^{\infty} p_k(t) z^k, \ p^*(z,s) = \int_0^{\infty} e^{-st} p(z,t), p_0^*(s) = \int_0^{\infty} e^{-st} p_0(t) dt.$$

Multiplying z^k on both sides of the previous equations and summing up, we obtain

$$\sum_{k=1}^{\infty} \frac{dp_k(t)}{dt} z^k = -(\lambda + \mu) \sum_{k=1}^{\infty} p_k(t) z^k + \lambda \sum_{k=1}^{\infty} p_{k-1}(t) z^k + \mu \sum_{k=1}^{\infty} p_{k+1}(t) z^k$$

$$\frac{\partial[p(z,t)-p_0(t)]}{\partial t} = -(\lambda+\mu)[p(z,t)-p_0(t)] + \lambda z p(z,t) + \frac{\mu}{z}[p(z,t)-p_0(t)-zp_1(t)]$$
$$\frac{\partial p(z,t)}{\partial t} = -\lambda p(z,t) - \mu[p(z,t)-p_0(t)] + \lambda z p(z,t) + \frac{\mu}{z}[p(z,t)-p_0(t)]$$
$$z\frac{\partial p(z,t)}{\partial t} = (1-z)[(\mu-\lambda z)p(z,t) - \mu p_0(t)]$$

Taking Laplace transform non both sides, we obtain

$$p^*(z,s) = \frac{zp(z,0) - \mu(1-z)p_0^*(s)}{sz-(1-z)(\mu-\lambda z)}.$$

4.7 Birth-Death Processes in Equilibrium

For most system design and analysis, what we are interested in is the system characterization in the steady-state (equilibrium, long-run). Let

$$p_k = \lim_{t\to\infty} p_k(t), \ \lim_{t\to\infty}\frac{dp_k(t)}{dt} = 0$$

where p_k is called *long-run probability* of finding the system with population k, or the stationary probability distribution.

From the B-D Chapman-Kolmogorov equations, we have

$$\begin{aligned} 0 &= -(\lambda_k+\mu_k)p_k + \lambda_{k-1}p_{k-1} + \mu_{k+1}p_{k+1}, \ k \geq 1 \\ 0 &= -\lambda_0 p_0 + \mu_1 p_1 \end{aligned}$$

From which we obtain

$$(\lambda_k+\mu_k)p_k = \lambda_{k-1}p_{k-1} + \mu_{k+1}p_{k=1}$$

which states that the flow rate out of state k is equal to the flow rate into state k — the balance equations!

From the balance equation, we have

$$p_{k+1} = \frac{1}{\mu_{k+1}}[(\lambda_k+\mu_k)p_k - \lambda_{k-1}p_{k-1}], p_1 = \frac{\lambda_0}{\mu_1}p_0,$$

which is a recursive formula. From this recursive formula, we could obtain the following:

$$p_k = \frac{\lambda_{k-1}}{\mu_k}p_{k-1},$$

i.e.,

$$\lambda_{k-1}p_{k-1} = \mu_k p_k,$$

which is called the *detailed balance equations*! The flow is directionally balanced: any state is memoryless, flow from one side will flow out on the other side. This property holds for many practical systems!

From the normalization equation: $\sum_{k=0}^{\infty} p_k = 1$, we obtain

$$p_0 = \left[1 + \sum_{k=1}^{\infty} \frac{\lambda_{k-1} \cdots \lambda_0}{\mu_k \cdots \mu_1}\right]^{-1}.$$

Therefore,

$$p_k = \frac{\frac{\lambda_{k-1} \cdots \lambda_0}{\mu_k \cdots \mu_1}}{1 + \sum_{k=1}^{\infty} \frac{\lambda_{k-1} \cdots \lambda_0}{\mu_k \cdots \mu_1}}, \quad k \geq 0.$$

Conditions for Stability: the conditions for the existence of steady-state (equilibrium).

Define

$$S_1 = \sum_{k=0}^{\infty} \left(\prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}}\right), \; S_2 = \sum_{k=0}^{\infty} \left(\frac{1}{\lambda_k \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}}}\right).$$

- All states of B-D process are *ergodic* iff $S_1 < \infty$ and $S_2 = \infty$
- All states are *recurrent null* iff $S_1 = \infty$ and $S_2 = \infty$
- All states are *transient* iff $S_1 = \infty$ and $S_2 < \infty$

If all states are *ergodic*, then the B-D process is stable, i.e., there exists a stationary probability distribution.

In particular, if there exists a δ satisfying $0 < \delta < 1$ such that $\frac{\lambda_k}{\mu_k} \leq \delta$, then the B-D is ergodic (there is a mistake in Kleinrock's book on this condition).

A special case: $\lambda_k = \lambda$ and $\mu_k = \mu$:

In this case, let $\rho = \lambda/\mu$. If $\rho < 1$, then the B-D process is ergodic, hence is stable, with

$$p_0 = 1 - \rho, \;\; p_k = (1 - \rho)\rho^k.$$

5 $M/M/1$ and its Variants

5.1 $M/M/1$

The $M/M/1$ queueing system consists of a single queueing station with single server, the arrivals form a Poisson process while the service distribution is exponential.

Let $N(t)$ denote the number of customers in the system at time t, $A(t)$ denotes the total number of arrivals from zero to time t. From the fact that $A(t_2)-A(t_1)$, the total arrivals in any time interval $[t_1, t_2]$ follows Poisson distribution with parameter $\lambda(t_2 - t_1)$, we have

$$\begin{aligned}
B_1: \quad & P(\text{exactly one arrival in } (t, t+\Delta t)|N(t)=k) \\
& = P(N(t+\Delta t)=k+1|N(t)=k) = P(A(t+\Delta t)-A(t)=1) \\
& = (\lambda\Delta t)e^{-\lambda\Delta t} = \lambda\Delta t + o(\Delta t) \\
D_1: \quad & P(\text{exactly one departure in } (t, t+\Delta t)|N(t)=k) \\
& P(N(t+\Delta t)=k-1|N(t)=k) = P(r(t)\le\Delta t) \ (r(t) \text{ is residual service time}) \\
& = 1-e^{-\mu\Delta t} = \mu\Delta t + o(\Delta t) \\
B_2: \quad & P(\text{no arrivals in } (t, t+\Delta t)|N(t)=k) = 1-\lambda\Delta t + o(\Delta t) \\
D_2: \quad & P(\text{no departure in } (t, t+\Delta t]|N(t)=k) = 1-\mu\Delta t + o(\Delta t)
\end{aligned}$$

Thus, $M/M/1$ is a B-D process. Hence, we have

$$\begin{aligned}
p_0 &= 1-\rho \\
p_k &= (1-\rho)\rho^k, \ \ \rho = \frac{\lambda}{\mu}
\end{aligned} \tag{1}$$

Moreover, the system is stable iff $\rho < 1$ or $\lambda < \mu$. We also have

$$N = \lim_{t\to\infty} = \sum_{n=0}^{\infty} np_n = \frac{\rho}{1-\rho} = \frac{\lambda}{\mu-\lambda}.$$

From Little's law, we obtain the *system time*

$$T = \frac{N}{\lambda} = \frac{\rho}{\lambda(1-\rho)} = \frac{1}{\mu-\lambda}.$$

The *waiting time* is

$$W = \frac{1}{\mu-\lambda} - \frac{1}{\mu} = \frac{\rho}{\mu-\lambda}.$$

from Little's law again, we obtain the average number of customers in the queue is given by

$$N_Q = \lambda W = \frac{\rho^2}{1-\rho}.$$

Remark: $\rho = 1 - p_0$, where p_0 is the probability of having no customers in the system. The average number in service will be

$$N_s = 0 \times P(\text{no customer in service}) + 1 \times P(\text{a customer in service}) = P(\text{server is busy}) = \rho$$

. Hence,

$$N_Q = N - N_s = \frac{\lambda}{\mu - \lambda} - \rho = \frac{\rho^2}{1-\rho}.$$

5.2 $M/M/m$

An $M/M/m$ queueing system is one in which arrivals form Poisson process with m exponential servers.

Let $N(t)$ denote the number of customers in the system at time t, then it can be shown that $N(t)$ is a B-D process with the following parameters

$$\begin{aligned} \lambda_k &= \lambda, \ \ k \geq 0 \\ \mu_k &= k\mu, \ \ 0 \leq k \leq m \\ \mu_k &= m\mu, \ \ k > m \end{aligned}$$

When $k \leq m$, we have

$$p_k = p_0 \prod_{i=0}^{k-1} \frac{\lambda_i}{\mu_{i+1}} = p_0 \prod_{i=0}^{k-1} \frac{\lambda}{(i+1)\mu} = p_0 \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!}.$$

We can obtain the similar result for the case when $k > m$. In summary, we obtain

$$p_k = \begin{cases} p_0 \frac{(m\rho)^k}{k!}, & k \leq m \\ p_0 \frac{m^m(\rho)^k}{m!}, & k > m \end{cases}$$

where

$$\rho = \frac{\lambda}{m\mu}.$$

From the normalization equation, we have

$$p_0 = \left[\sum_{k=0}^{m-1} \frac{(m\rho)^k}{k!} + \left(\frac{(m\rho)^m}{m!}\right)\left(\frac{1}{1-\rho}\right) \right]^{-1}.$$

The probability that an arriving customer is forced to join the queue is given by

$$p(\text{queueing}) = \sum_{k=m}^{\infty} p_k.$$

Thus, we can obtain

$$P_Q = P(\text{queueing}) = \frac{\left(\frac{(m\rho)^m}{m!}\right)\left(\frac{1}{1-\rho}\right)}{\left[\sum_{k=0}^{m-1}\frac{(m\rho)^k}{k!} + \left(\frac{(m\rho)^m}{m!}\right)\left(\frac{1}{1-\rho}\right)\right]}.$$

This is called *Erlang's C formula*.

Some other quantities are of interest to us. It is easy to derive the expected number of customers waiting in queue (not in service) is given by

$$N_Q = \sum_{n=0}^{\infty} n p_{n+m} = P_Q \frac{\rho}{1-\rho}.$$

The waiting time is given by

$$W = \frac{N_Q}{\lambda} = \frac{\rho P_Q}{\lambda(1-\rho)}.$$

Similarly, the average delay (system time) and the average number of customers in the system can be found very easily by repeatedly applying Little's Law.

5.3 $M/M/\infty$

This is the limiting case of $M/M/m$ where $m = \infty$. The detailed balance equation becomes

$$\lambda p_{k-1} = (k\mu) p_n,$$

from which we obtain

$$p_k = p_0 \left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!}$$

where

$$p_0 = \left[1 + \sum_{k=1}^{\infty}\left(\frac{\lambda}{\mu}\right)^k \frac{1}{k!}\right]^{-1} = e^{-\lambda/\mu}.$$

Thus,

$$p_k = \frac{(\lambda/\mu)^k}{k!} e^{-\lambda/\mu}, \;\; k = 0, 1, 2, \ldots,$$

which is a Poisson distribution! (Consider the number of people traveling in a city!) The average number of customers in the system and the system time are given by

$$N = \frac{\lambda}{\mu}, \; T = \frac{1}{\mu}.$$

Try to explain these results intuitively!

5.4 $M/M/m/m$

This is a system where the arrivals form a Poisson process, there are m exponential server, with no storage room (bufferless). This system is also called $m-$*server loss system*, which arises from telephone systems: a new arrival will be blocked (clear from the system) if all m servers are busy. The most important quantity is *blocking probability*.

The rate diagram can be drawn very easily. The detailed balance equations are

$$\lambda p_{k-1} = (k\mu)p_k, \;\; k = 0, 1, 2, \ldots, m.$$

We can easily find

$$p_k = \frac{(\lambda/\mu)^k/k!}{\sum_{i=0}^{m}(\lambda/\mu)^i/i!}, \; k = 0, 1, 2, \ldots, m.$$

Thus, the blocking probability is given by

$$p_m = \frac{(\lambda/\mu)^m/m!}{\sum_{i=0}^{m}(\lambda/\mu)^i/i!},$$

which is called *Erlang B formula*. In fact, this formula is also true for $M/G/m/m$ system!

Exercise: Find the average number of customers in the system and the system time.

Homework # 2 (from textbook)

3.1, 3.5, 3.6, 3.8, 3.9, 3.10, 3.12, 3.17, 3.19, 3.21.

6 $M/G/1$, $G/M/1$ and Priority Queues

6.1 $M/G/1$

This is the queueing system where the arrivals form a Poisson process and service distribution is generally distributed. The most important quantity we are interested in is the average waiting time.

Let X_i denote the service time for the $i-$th customer, assume that $\{X_1, X_2, \ldots\}$ are iid (independent and identically distributed). Let

$$\begin{aligned}
\overline{X} &= E[X] = \frac{1}{\mu} \\
\overline{X^2} &= E[X^2] = \text{second moment of service time} \\
W &-- \text{the average waiting time} \\
T &-- \text{the average system time} \\
\rho &= \lambda E[X] = \frac{\lambda}{\mu} - -\text{traffic intensity} \\
W_i &= \text{waiting time in queue of the } i-\text{th customer} \\
R_i &= \text{Residual service time seen by the } i-\text{th customer} \\
X_i &= \text{Service time of the } i-\text{th customer} \\
N_i &= \text{\# of customers found waiting in the queue by the } i-\text{th customer upon arrival}
\end{aligned}$$

Base on such notation, we obtain

$$W_i = R_i + \sum_{j=i-N_i}^{i-1} X_j.$$

By taking expectations and using the independence of the random variables N_i and $X_{i-1}, \ldots, X_{i-N_i}$, we have

$$E[W_i] = E[R_i] + E\left\{E\left\{\sum_{j=i-N_i}^{i-1} X_j \Big| N_i\right\}\right\} = E[R_i] + \overline{X}E[N_i],$$

where we have used the Wald's equation

$$E[\sum_{j=1}^{K} Y_i] = E[Y_i]E[K].$$

Taking limit $i \to \infty$ with some type of ergodicity assumption (such as $\lambda < \mu$), we obtain

$$W = R + \frac{1}{\mu} N_Q,$$

where

$$R = \text{mean residual time} = \lim_{i\to\infty} E[R_i].$$

By Little's law, we have $N_Q = \lambda W$, thus we have

$$W = R + \frac{1}{\mu}\lambda W = R + \rho W$$

hence

$$W = \frac{R}{1-\rho}.$$

Thus, the problem is reduced to finding R. Two approaches are presented here, we want to show that

$$R = \frac{1}{2}\lambda\overline{X^2}.$$

Method 1: In this method, we want to use the following Residual Life Theorem (Kleinrock's book or any other queueing books): Let Y_i denote iid random variables with distribution function $F(y)$ with average $1/\eta$, let r denote the residual life in generic form, then the Residual Life Theorem states that the pdf for r is given by $\eta[1-F(y)]$. Thus,

$$E[r] = \int_0^\infty y[\eta(1-F(y))dy = \frac{\eta}{2}(1-F(y))\Big|_0^\infty + \frac{\eta}{2}\int_0^\infty y^2 dF(y) = \frac{1}{2}\eta E[Y^2].$$

Now applying this result to our case: replacing Y_i by X_i, we have

$$R_i = \begin{cases} r_i & \text{server is busy} \\ 0 & \text{server is idle} \end{cases}$$

where r_i is the residual life found by the $i-$th customer which has the stationary distribution of generic form r. We thus obtain

$$R = E[R_i] = 0 \times P(\text{idle}) + E[r_i]P(\text{busy}) = \rho E[r] = \rho \times \frac{1}{2}\mu E[X^2] = \frac{\lambda}{2}\overline{X^2}.$$

Method 2 (graphical or Takacs's technique) We can plot the residual service time $r(\tau)$ (shown in Figure 2): notice that when a new service of the customer X begins, $r(\tau)$ starts at X and decays linearly for X time units (with slope -1). Consider a time for which $r(t) = 0$. The time average of $r(t)$ in the interval $[0, t]$ is

$$\frac{1}{t}\int_0^t r(\tau)d\tau = \frac{1}{t}\sum_{i=1}^{M(t)}\frac{1}{2}X_i^2 = \frac{1}{t}\text{the area under curve } r(\tau)$$

$$\frac{1}{2}\frac{M(t)}{t}\frac{\sum_{i=1}^{M(t)}X_i^2}{M(t)}$$

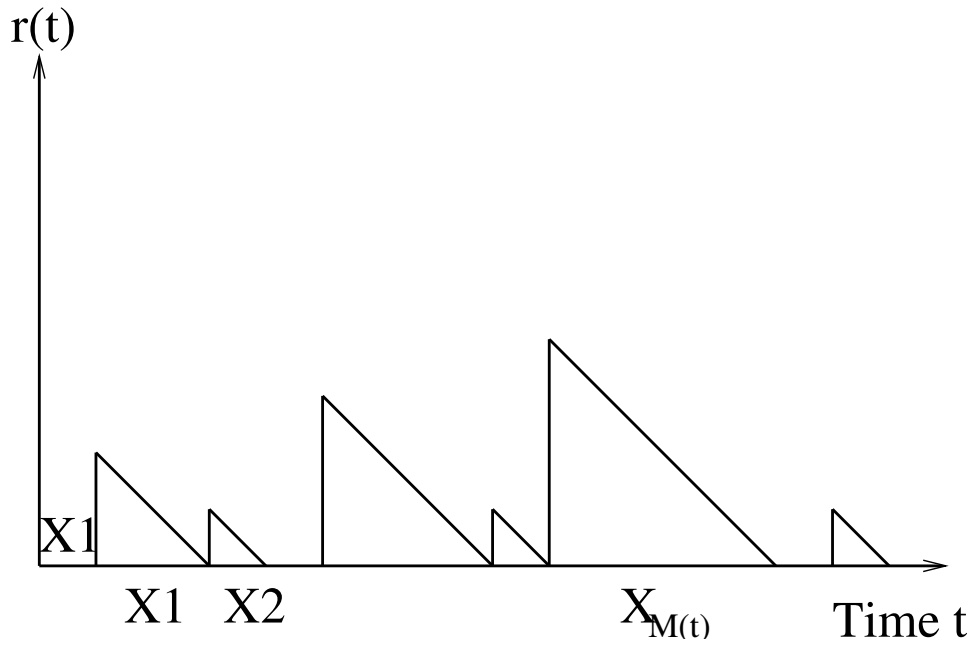


Figure 2: Derivation of the average residual service time

where $M(t)$ is the number of service completions within $[0, t]$. Assuming all limits exist, taking limit $t \to \infty$, we obtain

$$R = \frac{1}{t}\int_0^t r(\tau)d\tau = \frac{1}{2}\lim_{t\to\infty}\frac{M(t)}{t}\lim_{t\to\infty}\frac{\sum_{i=1}^{M(t)} X_i^2}{M(t)} = \frac{1}{2}\lambda\overline{X^2}.$$

Final results: We finally arrive at

$$W = \frac{\lambda\overline{X^2}}{2(1-\rho)}.$$

This is called *Pollaczek-Khinchin* formula or simply *P-K formula*.

Using Little's law, we have the following results

$$\begin{aligned} T &= \overline{X} + \frac{\lambda\overline{X^2}}{2(1-\rho)} \\ N_Q &= \frac{\lambda^2\overline{X^2}}{2(1-\rho)} \\ N &= \rho + \frac{\lambda^2\overline{X^2}}{2(1-\rho)} \end{aligned}$$

Remark: Indeed, for $M/G/1$ queue, all average values for the queueing system depend on the second moment statistics! Thus, using $M/M/1$ approximation may NOT be a good idea.

Special cases:

$$\begin{aligned} W &= \frac{\rho}{\mu(1-\rho)} \quad (M/M/1) \\ W &= \frac{\rho}{2\mu(1-\rho)} \quad (M/D/1) \end{aligned}$$

Exercises:

(a). For $M/D/1$ system, derive the formulae for the average number of customers in the system, in the queue and the average system time.

(b). Show that the $M/D/1$ gives the minimum waiting time among all $M/G/1$ queues given the same arrival rate and service rate.

6.2 More General Results for $M/G/1$

What if I want to find information about the queueing delay? More rigorous machinery is needed. The following material is from Kleinrock's book.

6.2.1 Imbedded Markov Chain

To analyze this queue, we need to find the imbedded Markov chain. For $M/G/1$, the number of customers in the queue, say, $N(t)$ is no longer a Markov chain because the current state $N(t) = k$ is not enough to characterize the future: since the service time is no longer memoryless, the customer in service is "different" from those in queue in terms of service time distribution from now to the ending. However, to make them the same, we could observe the number of customers observed in some special instants. For example, in this case, at the end of service completion (the departure instants), the observed number of customers in the system (excluding the finished customers) will all have the same service times statistically, thus it could be shown that such "sampled" sequence will form a Markov chain from which we could derive all interested quantity. This leads to the so-called imbedded Markov chain. If we define the following quantities,

$$\begin{aligned} p_k &= \lim_{t\to\infty} P(N(t) = k) = P[\text{there are k customers in the system at time t}) \\ r_k &= \lim_{t\to\infty} P[\text{arrival at time t finds k customers in the system}) \\ d_k &= \lim_{t\to\infty} P[\text{departure at time t finds k customers in the system}) \end{aligned}$$

it can be shown that for $M/G/1$, $p_k = r_k = d_k$.

6.2.2 Transition Probabilities

Define

$$\begin{aligned}
C_n &= \text{the nth customer to enter the system} \\
t_n &= \text{the arrival time instant of } C_n \\
\tau_n &= t_n - t_{n-1} \\
X_n &= \text{service time for } C_n \\
q_n &= \text{number of customers left behind by departure of } C_n \text{ from service} \\
v_n &= \text{number of customers arriving during the service of } C_n
\end{aligned}$$

It can be easily shown that $\{q_n\}$ is a Markov chain and we can find the probability distribution of this chain in order to find all other quantities. Intuitively speaking, this Markov chain is homogeneous, and thus we just offer the stationary distribution for illustration purpose. Let $p_{ij} = P(q_{n+1} = j | q_n = i)$, and $\alpha_k = P(v_{n+1} = k)$. Since we only have one single server, thus the state decrements at most 1. However, since we can have multiple arrivals during a service time, and thus the transition to higher states are possible. It can be easily found that the transition matrix is given by

$$\begin{pmatrix}
\alpha_0 & \alpha_1 & \alpha_2 & \alpha_3 & \cdots \\
\alpha_0 & \alpha_1 & \alpha_2 & \alpha_3 & \cdots \\
0 & \alpha_0 & \alpha_1 & \alpha_2 & \cdots \\
0 & 0 & \alpha_0 & \alpha_1 & \cdots \\
0 & 0 & 0 & \alpha_0 & \cdots \\
\cdot & \cdot & \cdot & \cdot & \cdots \\
\cdot & \cdot & \cdot & \cdot & \cdots
\end{pmatrix}$$

The parameter α_k can be easily computed as follows: Let $b(x)$ denote the probability density function (pdf) for the service time, then we have

$$\begin{aligned}
\alpha_k &= P(v_{n+1} = k) = \int_0^\infty P(v_{n+1} = k | X_{n+1} = x) b(x) dx \\
&= \int_0^\infty \frac{(\lambda x)^k}{k!} e^{-\lambda x} b(x) dx
\end{aligned}$$

If we let $V(z)$ denote the moment generating function for v_k and let $b^*(s)$ denote the Laplace transform for $b(x)$, then we can easily obtain

$$\begin{aligned}
V(z) &= \sum_{k=0}^\infty P(v_{n+1} = k) z^k = \int_0^\infty e^{-\lambda x} \left(\sum_{k=0}^\infty \frac{(\lambda x z)^k}{k!} \right) b(x) dx \\
&= \int_0^\infty e^{-\lambda x} e^{\lambda x z} b(x) dx = \int_0^\infty e^{-(\lambda - \lambda z)x} b(x) dx = b^*(\lambda - \lambda z). \qquad (2)
\end{aligned}$$

6.2.3 Average Queue Length

In fact, there is a close relationship between the queue length and number of arrivals during service time. It is easy to verify that when $q_n > 0$, the customer just finished service is C_{n+1}, while q_n includes this one when C_n departed. Therefore, when C_{n+1} completes the service and counts the number of customers excluding itself would be $(q_n - 1)$ plus the number of new arrivals during its service, and thus $q_{n+1} = q_n - 1 + v_{n+1}$. When $q_n = 0$, only when C_{n+1} arrives and can receive immediately receiving service. When it leaves, only the number of new arrivals during its service is observed, and thus $q_{n+1} = v_{n+1}$. Define

$$\Delta_k = \begin{cases} 1 & k > 0 \\ 0 & k \leq 0 \end{cases}$$

Then, we obtain the fundamental equation for queue length

$$q_{n+1} = q_n - \Delta_{q_n} + v_{n+1} \tag{3}$$

It is easy to obtain the following (the overline denotes the expectation operator as before)

$$\overline{\Delta_{q_n}} = \sum_{k=0}^{\infty} \Delta_k P(q_n = k) = \sum_{k=1}^{\infty} P(q_n = k) = P(q_n > 0) = P(\text{busy system}) = \rho$$

where $\rho = \lambda \overline{X}$.

By taking average on both sides of equation (3) and letting $\tilde{v}$ denote the generic form (stationary version) of v_n, we obtain

$$\overline{\tilde{v}} = \overline{\Delta_{q_n}} = \rho.$$

Now, squaring both sides of equation (refeq:ql) and also noticing that $(\Delta_{q_n})^2 = \Delta_{q_n}$ and $q_n \Delta_{q_n} = q_n$, we obtain

$$q_{n+1}^2 = q_n^2 + \Delta_{q_n}^2 + v_{n+1}^2 - 2q_n\Delta_{q_n} + 2q_n v_{n+1} - 2\Delta_{q_n} v_{n+1}$$

i.e.,

$$q_{n+1}^2 = q_n^2 + \Delta_{q_n} + v_{n+1}^2 - 2q_n + 2q_n v_{n+1} - 2\Delta_{q_n} v_{n+1}$$

By taking the expectation on both sides, we obtain

$$\overline{q_{n+1}^2} = \overline{q_n^2} + \overline{\Delta_{q_n}} + \overline{v_{n+1}^2} - 2\overline{q_n} + 2\overline{q_n v_{n+1}} - 2\overline{\Delta_{q_n} v_{n+1}}$$

By noticing that q_n and v_{n+1} are independent, we obtain

$$\overline{q_{n+1}^2} = \overline{q_n^2} + \overline{\Delta_{q_n}} + \overline{v_{n+1}^2} - 2\overline{q_n} + 2\overline{q_n} \cdot \overline{v_{n+1}} - 2\overline{\Delta_{q_n}} \cdot \overline{v_{n+1}}$$

Letting n go to ∞, we obtain

$$0 = \overline{\Delta_{\tilde{q}}} + \overline{\tilde{v}^2} - 2\bar{\tilde{q}} + 2\bar{\tilde{q}} \cdot \bar{\tilde{v}} - 2\overline{\Delta_{\tilde{q}}} \cdot \bar{\tilde{v}}$$

or,

$$0 = \rho + \overline{\tilde{v}^2} - 2\bar{\tilde{q}} + 2\bar{\tilde{q}} \cdot \bar{\tilde{v}} - 2\rho\bar{\tilde{v}}$$

Thus, we have

$$\bar{\tilde{q}} = \rho + \frac{\overline{\tilde{v}^2} - \rho}{2(1-\rho)}.$$

From equation (2) and the relationship between the moments and moment generating function/Laplace transform, we can easily obtain

$$V''(1) = \overline{\tilde{v}^2} - \bar{\tilde{v}} = V''(z)|_{z=1} = b^{*\prime\prime}(\lambda - \lambda z)|_{z=1} = \lambda^2 b^{*\prime\prime}(0) = \lambda^2 \overline{X^2}.$$

Thus, we finally obtain the P-K equation for queue length as

$$\bar{\tilde{q}} = \rho + \frac{\lambda^2 \overline{X^2}}{2(1-\rho)} \tag{4}$$

Let T and W denote the average time and average waiting time as before. From Little's Law, we can easily obtain

$$\begin{aligned} T &= \frac{\bar{\tilde{q}}}{\lambda} = \frac{1}{\mu} + \frac{\lambda \overline{X^2}}{2(1-\rho)} \\ W &= T - 1/\mu = \frac{\lambda \overline{X^2}}{2(1-\rho)} \end{aligned}$$

which offers another derivation of P-K formula for average waiting time.

6.2.4 Distribution of the Number of Customers in the System

In fact, we could do better than that from the fundamental equation for queue length. Let $Q_n(z)$ denote the moment generating function of q_n, and $Q(z)$ denote the moment generating function of the generic form (i.e., the limiting case, $\lim_{n\to\infty} Q_n(z) = Q(z)$). From equation (3), we obtain (noting the independence of v_{n+1} and past history)

$$Q_{n+1}(z) = \overline{z^{q_{n+1}}} = \overline{z^{q_n - \Delta_{q_n} + v_{n+1}}} = \overline{z^{q_n - \Delta_{q_n}}} \cdot \overline{z^{v_{n+1}}} = V(z)\overline{z^{q_n - \Delta_{q_n}}} \tag{5}$$

Moreover, we have

$$\begin{aligned} \overline{z^{q_n - \Delta_{q_n}}} &= \sum_{k=0}^{\infty} P(q_n = k) z^{k - \Delta_k} \\ &= P(q_n = 0) + \sum_{k=1}^{\infty} P(q_n = k) z^{k-1} = P(q_n = 0) + \frac{1}{z} \sum_{k=1}^{\infty} P(q_n = k) z^k \\ &= P(q_n = 0) + \frac{1}{z}[Q_n(z) - P(q_n = 0)] \end{aligned}$$

Thus, we have

$$Q_{n+1}(z) = V(z)\left(P(q_n = 0) + \frac{Q_n(z) - P(q_n = 0)}{z}\right).$$

Letting $n \to \infty$, we obtain

$$Q(z) = V(z)\left(P(\tilde{q} = 0) + \frac{Q(z) - P(\tilde{q} = 0)}{z}\right).$$

By noticing that $P(\tilde{q} = 0) = 1 - \rho$, we have

$$Q(z) = V(z)\frac{(1-\rho)(1-1/z)}{1 - V(z)/z}.$$

Taking (2) into this equation, we finally obtain

$$Q(z) = b^*(\lambda - \lambda z)\frac{(1-\rho)(1-z)}{b^*(\lambda - \lambda z) - z} \tag{6}$$

which is called P-K transformation formula.

As the special case, $M/M/1$ has $b^*(s) = \mu/(s+\mu)$. Taking this into (6), we obtain

$$Q(z) = \frac{1-\rho}{1-\rho z}.$$

From this we can easily obtain that $P(\tilde{q} = k) = (1-\rho)\rho^k$.

6.2.5 Distribution of the Waiting Time

We can study this by first studying the system time. Let s_n denote the total time the customer C_n spent in the system, then we have $s_n = w_n + X_n$. If we examine the analogy between the v_{k+1} and X_{n+1}, we can think that during s_n time period, there are q_n arrivals by recalling that q_n is the number of customers left behind when nth customer departs (by equating q_n to v_{k+1} and s_n to X_{n+1}. With this argument, we can obtain equation similar to (2). If we use $s^*(s)$ denote the Laplace transform of the pdf of the generic service time, we have

$$Q(z) = s^*(\lambda - \lambda z) = b^*(\lambda - \lambda z)\frac{(1-\rho)(1-z)}{b^*(\lambda - \lambda z) - z} \tag{7}$$

From this we can easily obtain

$$s^*(s) = b^*(s)\frac{s(1-\rho)}{s - \lambda + \lambda b^*(s)} \tag{8}$$

Let $w^*(s)$ denote the Laplace transform of the pdf of the generic waiting time for the customer C_n. Since X_n is independent of the past history (e.g., w_n), we have

$$s*(s) = \overline{e^{-ss_n}} = \overline{e^{-sw_n}e^{-sX_n}} = \overline{e^{-sw_n}} \cdot \overline{e^{-sX_n}} = w^*(s)b^*(s).$$

Thus, we obtain

$$w^*(s) = \frac{s^*(s)}{b^*(s)} = \frac{s(1-\rho)}{s - \lambda + \lambda b^*(s)} \tag{9}$$

If we let

$$\hat{b}^*(s) = \frac{1 - b^*(s)}{s\overline{X}},$$

then we obtain a nicer result

$$w^*(s) = \frac{1-\rho}{1-\rho\hat{b}^*(s)} = (1-\rho)\sum_{k=0}^{\infty}\rho^k[\hat{b}^*(s)]^k \tag{10}$$

If we let $\hat{b}(x)$ denote the pdf corresponding to $\hat{b}^*(s)$ and $\hat{b}^{(k)}(x)$ denote the self-convolution of $\hat{b}(x)$ k times, then the pdf of waiting time is given by

$$w(x) = \sum_{k=0}^{\infty}(1-\rho)\rho^k\hat{b}^{(k)}(x) \tag{11}$$

For $M/M/1$, we can easily obtain

$$\begin{aligned} s(x) &= \mu(1-\rho)e^{-\mu(1-\rho)x},\ x \geq 0 \\ w(x) &= (1-\rho)\delta(x) + \lambda(1-\rho)e^{-\mu(1-\rho)x},\ x \geq 0 \end{aligned}$$

where $\delta(x)$ is an impulse function (Dirac function).

6.3 $G/M/1$

In order to give a clear presentation, we may have to introduce some additional notation and new techniques, due to the time limitation, we will not give the details in this course. However, the following is the result. The imbedded Markov chain is the "sampled" sequence at the arrival time instants, i.e., the number of customers in the system upon the arrival time instants. Detailed can be found in Kleinrock's book.

Let p_k denote the steady-state probability distribution for $G/M/1$, let $A(t)$ denote the probability density function of the interarrival time with the Laplace transform $A^*(s)$, let $1/\mu$ denote the average service time, then we have

$$p_k = (1-\sigma)\sigma^k,\ \ k = 0, 1, 2, \ldots$$

where σ is the solution of

$$\sigma = A^*(\mu(1-\sigma))$$

in the range $0 < \sigma < 1$.

With this distribution, we could find all other quantities of interest. Please find them as an exercise.

6.4 $G/G/1$

No exact analytical results are available up to now. However, an upper bound for the average waiting time can be obtained.

Assume that the inter-arrival time and service times are independent! (Very important assumption). Let

$$\begin{aligned}
\sigma_a^2 &= \text{variance of the interarrival times} \\
\sigma_b^2 &= \text{variance of the service times} \\
1/\lambda &= \text{average interarrival time} \\
1/\mu &= \text{average service time} \\
\rho &= \lambda/\mu - \text{traffic intensity factor} \\
W_k &= \text{waiting time of the } k\text{th customer} \\
X_k &= \text{service time of the } k\text{th customer} \\
\tau_{k+1} &= \text{interarrival time between the } k\text{th and } k+1\text{th customer}
\end{aligned}$$

Then we have the following obvious relationship

$$W_{k+1} = \max\{0, W_k + X_k - \tau_k\} = (W_k + X_k - \tau_k)^+$$

We use the following notation:

$$Y^+ = \max\{0, Y\},\ Y^- = -\min\{0, Y\} = \max\{0, -Y\},\ \overline{Y} = E(Y),\ \sigma_Y^2 = E[Y^2 - \bar{Y}^2].$$

We have

$$Y = Y^+ - Y^-,\ \ Y^+Y^- = 0$$

from which we obtain

$$\overline{Y} = \overline{Y^+} - \overline{Y^-},\ \sigma_Y^2 = \sigma_{Y^+}^2 + \sigma_{Y^-}^2 + 2\overline{Y^+} \cdot \overline{Y^-}.$$

Let

$$V_k = X_k - \tau_k,\ \ I_k = (W_k + V_k)^-,$$

we have

$$W_{k+1} = (W_k + V_k)^+.$$

Notice that I_k is the length of the idle period between the arrival of the kth customer and the arrival of the $(k+1)$th customer.

Thus, we have

$$\begin{aligned}\sigma^2_{W_k+V_k} &= \sigma^2_{(W_k+V_k)^+} + \sigma^2_{(W_k+V_k)^-} + 2\overline{(W_k+V_k)^+}\cdot\overline{(W_k+V_k)^-} \\ &= \sigma^2_{W_{k+1}} + \sigma^2_{I_k} + 2\overline{W_{k+1}}\cdot\overline{I_k}\end{aligned}$$

However, using the independence of W_k and V_k and the independence of X_k and τ_k, we also have

$$\begin{aligned}\sigma^2_{(W_k+V_k)} &= \sigma^2_{W_k} + \sigma^2_{V_k} \\ &= \sigma^2_{W_k} + \sigma^2_a + \sigma^2_b\end{aligned}$$

Combining the above two equations, we obtain

$$\sigma^2_{W_k} + \sigma^2_a + \sigma^2_b = \sigma^2_{W_{k+1}} + \sigma^2_{I_k} + 2\overline{W_{k+1}}\cdot\overline{I_k}.$$

Letting $k \to \infty$ and using

$$\overline{W_k} \to W,\ \sigma^2_{W_k} \to \sigma^2_W,\ \overline{I_k^2} \to I,\ \sigma^2_{I_k} \to \sigma^2_I,$$

we obtain

$$W = \frac{\sigma^2_a + \sigma^2_b}{2I} - \frac{\sigma^2_I}{2I}.$$

We notice that the idle period between consecutive arrivals will be the time accumulation of times that the server is idle, while the server is idle with probability $1-\rho$, thus we obtain

$$I = (1-\rho) * (1/\lambda).$$

Hence

$$W = \frac{\lambda(\sigma^2_a + \sigma^2_b)}{2(1-\rho)} - \frac{\lambda\sigma^2_I}{2(1-\rho)},$$

from which we finally arrive at

$$W \le \frac{\lambda(\sigma^2_a + \sigma^2_b)}{2(1-\rho)}.$$

As an exercise, compare the upper bound with the real value for W in $M/G/1$, $M/M/1$ and $M/D/1$.

6.5 Priority Queues

Prioritized queues become more important than ever nowadays, in particular in multimedia networks. We will concentrate on the non-preemptive priority, although preemptive queues (such as in CDPD or cognitive radios) are also important. We limit our discussion for $M/G/1$ systems.

Let λ_k denote the arrival rate for the class k customers among n different priority classes, assume that the class k customers have service time X_k of general distribution, some independence assumptions are used.

Let

$$\begin{aligned} N_Q^k &= \text{average number of customers in the queue for priority k} \\ W_k &= \text{average queueing time for priority k} \\ \rho_k &= \frac{\lambda_k}{\mu_k} = \text{traffic intensity for priority k} \\ R &= \text{mean residual service time} \end{aligned}$$

Assume that the overall system traffic intensity is less than unity:

$$\rho = \sum_{i=1}^{n} \rho_i < 1.$$

The Intuitive argument: For the highest priority, we have

$$W_1 = R + \frac{1}{\mu_1} N_Q^1$$

which indicates that the waiting time for class 1 should be equal to the average residual service time and the time to drain all waiting customers (here we use μ_1 rather than λ_1 !!)

From Little's law, we have

$$N_Q^1 = \lambda_1 W_1,$$

thus we have

$$W_1 = R + \frac{1}{\mu_1}(\lambda_1 W_1) = R + \rho_1 W_1,$$

hence

$$W_1 = \frac{R}{1-\rho_1}.$$

Similarly, for the class 2 customers, we have

$$W_2 = R + \frac{1}{\mu_1} N_Q^1 + \frac{1}{\mu_2} N_Q^2 + \frac{1}{\mu_1}(\lambda_1 W_2) = R + \rho_1 W_1 + \rho_2 W_2 + \rho_1 W_2$$

which can be interpreted as follows: the waiting time for a class 2 is equal to the residual service time, plus the time for all waiting class 1 and 2 customers to finish, plus the time for class 1 customers to arrive during the waiting time! Thus, we obtain

$$W_2 = \frac{R + \lambda_1 W_1}{1 - \rho_1 - \rho_2} = \frac{R}{(1-\rho_1)(1-\rho_1-\rho_2)}.$$

When $k > 2$, we can combine all customers into one class with traffic intensity $\tilde{\rho}_1 = \rho_1 + \rho_2 + \cdots + \rho_{k-1}$, applying the case when $k = 2$, we obtain

$$W_k = \frac{R}{(1-\tilde{\rho}_1)(1-\tilde{\rho}_1-\rho_k)} = \frac{R}{(1-\rho_1-\cdots-\rho_{k-1})(1-\rho_1-\cdots-\rho_k)}.$$

We can also show using a similar argument as for $k = 2$, a class k customer has to wait for all higher priority customers in the queue and arriving during the waiting time this customer in waiting, thus once again, we obtain (by noticing that $N_Q^i = \lambda_i W_i$)

$$W_k = R + \sum_{i=1}^{k-1} \frac{1}{\mu_i} N_Q^i + \frac{1}{\mu_k} N_Q^k + \sum_{i=1}^{k-1} \frac{1}{\mu_i}(\lambda_i W_2) = R + \sum_{i=1}^{k-1} \rho_i W_i + (\sum_{i=1}^{k} \rho_i) W_2$$

from which we obtain

$$W_k = \frac{R}{(1-\rho_1-\cdots-\rho_{k-1})(1-\rho_1-\cdots-\rho_k)}.$$

Here we use the identity

$$1+\frac{\rho_1}{1-\rho_1}+\frac{\rho_2}{(1-\rho_1)(1-\rho_1-\rho_2)}+\cdots+\frac{\rho_k}{(1-\rho_1-\cdots-\rho_{k-1})(1-\rho_1-\cdots-\rho_k)} = \frac{1}{1-\rho_1-\cdots-\rho_k}.$$

Thus, the problem is reduced to finding R.

Method 1 for finding R: Let r denote the residual service time in generic form, let $\lambda = \sum_{i=1}^n \lambda_i$, then we have

$$\begin{aligned} R &= E[r] = \sum_{i=1}^{n} E[r|\text{class i customer in service}]P(\text{class i customer in service}) \\ &= \sum_{i=1}^{n} \left(\frac{1}{2}\lambda\overline{X_i^2}\right)\left(\frac{\lambda_i}{\lambda}\right) = \frac{1}{2}\sum_{i=1}^{n}\lambda_i\overline{X_i^2} \end{aligned}$$

Method 2 for finding R: Let $M_i(t)$ be the number of class i customers arriving during $[0,t]$ and $X_{i,j}$ denote the service time for the $j-$th customer belonging to class i, we have

$$\begin{aligned} R &= \lim_{t\to\infty} \frac{1}{t} \sum_{i=1}^{n} \sum_{j=1}^{M_i(t)} \frac{1}{2} X_{i,j}^2 \\ &= \sum_{i=1}^{n} \lim_{t\to\infty} \left(\frac{M_i(t)}{t}\right)\left(\frac{\sum_{j=1}^{M_i(t)} \frac{1}{2} X_{i,j}^2}{M_i(t)}\right) \\ &= \frac{1}{2}\sum_{i=1}^{n}\lambda_i\overline{X_i^2}. \end{aligned}$$

In sum, we finally obtain

$$W_k = \frac{\sum_{i=1}^n \lambda_i \overline{X_i^2}}{2(1-\rho_1-\cdots-\rho_{k-1})(1-\rho_1-\cdots-\rho_k)}.$$

Homework #3

3.30, 3.31, 3.36, 3.37

7 Time-Reversibility and Multidimensional Markov Chain

7.1 Time-Reversibility: Burke's Theorem

Birth-death processes have been used for all studies in $M/M/1$, $M/M/m$, $M/M/\infty$, and $M/M/m/m$, the so-called detailed balance equations are given by

$$\begin{aligned} p_j p_{j(j+1)} &= p_{j+1} p_{(j+1)j} \text{ (discrete-time MC)} \\ p_j q_{j(j+1)} &= p_{j+1} q_{(j+1)j} \text{ (continuous-time MC)} \end{aligned}$$

Time-reversibility is the generalized property of the detailed balance equations: a process will be the same no matter whether we look forward or backward in time! To demonstrate, consider an irreducible, discrete-time Markov chain $\{X_n\}$ having transition probabilities p_{ij} and stationary distribution $\{p_j\}$: $p_{[}j = \Pr\{X_n = j\}$ in steady-state.

Suppose that we look at the MC backward in time, $\cdots, X_n, X_{n-1}, \cdots$, then the future in the forward chain becomes the past in the backward chain, thus we have

$$\begin{aligned} &\Pr(X_m = j | X_{m+1} = i, X_{m+2} = i_2, \ldots, X_{m+k} = i_k) \\ &= \frac{\Pr(X_m = j, X_{m+1} = i, X_{m+2} = i_2, \ldots, X_{m+k} = i_k)}{\Pr(X_{m+1} = i, X_{m+2} = i_2, \ldots, X_{m+k} = i_k)} \\ &= \frac{\Pr(X_m = j, X_{m+1} = i) \Pr(X_{m+2} = i_2, \ldots, X_{m+k} = i_k | X_m = j, X_{m+1} = i)}{\Pr(X_{m+1} = i) \Pr(X_{m+2} = i_2, \ldots, X_{m+k} = i_k | X_{m+1} = i)} \\ &= \frac{\Pr(X_m = j, X_{m+1} = i)}{\Pr(X_{m+1} = i)} = \frac{\Pr(X_m = j) \Pr(X_{m+1} = i | X_m = j)}{\Pr(X_{m+1} = i)} = \frac{p_j p_{ji}}{p_i} \end{aligned}$$

This implies that the backward process is also a Markov process with transition probability

$$p^*_{ij} = \Pr(X_m = j | X_{m+1} = i) = \frac{p_j p_{ji}}{p_i}.$$

Definition: If $p^*_{ij} = p_{ij}$, i.e., the transition probabilities of the forward and the reserved chain are identical, then the chain is called *time reversible*.

Properties: For irreducible, aperiodic and time-reversible MC, we have

(1). The reversed MC is also irreducible, aperiodic and has the same stationary distribution as the forward MC;

(2). If we can find positive numbers p_i satisfying $\sum_{i=0}^{\infty} p_i = 1$ such that

$$p_{ij}^* = \frac{p_j p_{ji}}{p_i}, \;\; i, j \geq 0$$

forms a probability transition matrix, i.e., $\sum_{j=0}^{\infty} p_{ij}^* = 1$ for $i \geq 0$, then $\{p_i\}$ is the stationary distribution and p_{ij}^* are the transition probabilities of the reversed MC. *Important observation*: This property holds *regardless of whether the chain is reversible or not*!

(3). A chain is time reversible if and only if the detailed balance equations hold: $p_i p_{ij} = p_j p_{ji}$ for any $i, j \geq 0$. This implies that B-D processes, $M/M/1$, $M/M/\infty$ are all time-reversible!

Continuous-time MC

For continuous-time MC $X(t)$, the process $X(-t)$ is the reversed chain.

Theorem: For the continuous-time MC having transition rates q_{ij} and stationary distribution $p_j > 0$ for $j \geq 0$, we have

(1) The reversed chain $X(-t)$ is also a continuous-time MC with the same stationary distribution and with the transition rates given by

$$q_{ij}^* = \frac{p_j q_{ji}}{p_i}, \; i, j \geq 0$$

(2) If we find positive numbers p_j with $\sum_{i=0}^{\infty} p_i = 1$ such that

$$q_{ij}^* = \frac{p_j q_{ji}}{p_i}, \; i, j \geq 0$$

and satisfy

$$\sum_{j=0}^{\infty} q_{ij} = \sum_{j=0}^{\infty} q_{ij}^* \;\; \text{(total balance equation)}$$

then $\{p_i\}$ is the stationary distribution of the both forward and backward chains, and q_{ij}^* are the transition rates of the reversed chain.

(3) The forward chain is time-reversible if and only if the detailed balance equations hold:

$$p_i q_{ij} = p_j q_{ji}, \;\; i, j \geq 0$$

Remark: Total balance equations are

$$p_i \sum_{j=0}^{\infty} q_{ij} = \sum_{j=0}^{\infty} p_j p_{ji}, \; i \geq 0.$$

Burke's Theorem: Consider an $M/M/1$, $M/M/\infty$ or $M/M/m$ system with arrival rate λ. Suppose that the system starts in steady-state (i.e., the initial probability distribution is the stationary distribution), then the following statements are true:

(a). The departure process is also Poisson with rate λ;

(b). At each time t, the number of customers in the system is independent of the sequence of departure times prior to t.

Burke's Theorem (1956): In an $M/G/1$ system, the departure process is Poisson if and only if the service time is exponentially distributed. I.e., $M/M/1$ is the ONLY $M/G/1$ that the departure process is Poisson!

Example: Tandem queues (two queue systems in tandem). Assume that the arrival process and the service times are all independent. If we assume that the arrival process is Poisson and all service times for both queues are exponentially distributed, then from Burke's Theorem, the departure process of the first queue is also Poisson, hence the second queue in tandem is also $M/M/1$, hence this tandem is in fact a tandem of $M/M/1$ queues. Assuming that the arrival rate is λ, and the average service times are $1/\mu_1$ and $1/\mu_2$ respectively. Let $\rho_1 = \lambda/\mu_1$ and $\rho_2 = \lambda/\mu_2$ and $p(n_1, n_2)$ denote the probability distribution that there are n_1 customers in the first queue and n_2 customers in the second queue. Then, we have

$$p(n_1, n_2) = \rho_1^{n_1}(1-\rho_1)\rho_2^{n_2}(1-\rho_2).$$

This is a network of queues!

7.2 Multidimensional Markov Chain (MMC)

MMC arises in many situations:

- multiple classes of customers
- multiple priority classes
- multimedia networks
- QoS service networks

Consider a queueing system with K customer types, the process–# of customers in the system at time t, N(t)– can only be represented as

$$N(t) = (n_1(t), n_2(t), \ldots, n_K(t))$$

where $n_i(t)$ is the number of customers of type i at time t.

Example 1: Two traffic streams of different arrival rates merge into one stream for service, if the average service times for two traffic streams are the same, then we can treat all customers the same with the aggregated rate (the sum of the rates of the two), the process will be the total number of customers in the system. However, if two streams have different average service times, then the total number of customers in the system will not tell the whole story (or does not convey the whole information), because customers from different streams will demand different services! The only way is to use the two dimensional process to accurately capture the dynamic of the queueing system.

Example 2: A cellular system: each cell can be modeled as a queueing system, two traffic streams will request for services, the new calls and handoff calls, usually the handoff call will need preferential treatment, the channel holding times (service times) will be different fro new calls and handoff calls.

Let $(n_1, n_2, \ldots, n_K)$ denote a state, $p(n_1, n_2, \ldots, n_K)$ denote the steady-state probability distribution. Let

$$n(j+) = (n_1, n_2, \ldots, n_{j-1}, n_j + 1, n_{j+1}, \ldots, n_K)$$
$$n(j-) = (n_1, n_2, \ldots, n_{j-1}, n_j - 1, n_{j+1}, \ldots, n_K)$$

Similar to the case for one-dimensional MC, we only need to observe the state transition by one unit. This leads to the B-D process:

$$n \longrightarrow n(j+): \quad \text{one type-j customer arrives}$$
$$n \longrightarrow n(j-): \quad \text{one type-j customer departs}$$

If we could find $p(n) = p(n_1, n_2, \ldots, n_K)$, which satisfies the "detailed balance" equations:

$$\lambda_j p(n_1, n_2, \ldots, n_{j-1}, n_j, n_{j+1}, \ldots, n_K) = \mu_j p(n_1, n_2, \ldots, n_{j-1}, n_j + 1, n_{j+1}, \ldots, n_K)$$

then $p(n)$ will be the stationary distribution. If

$$p(n_1, n_2, \ldots, n_{j-1}, n_j, n_{j+1}, \ldots, n_K) = p_1(n_1)p_2(n_2)\cdots p_K(n_k),$$

then the queueing system is said to possess *product form*.

Truncation of independent multiple single-class systems

Consider K independent $M/M/1$ queues: state $n = (n_1, n_2, \ldots, n_K)$, for each queue i, the distribution is

$$p_i(n_i) = \rho_i^{n_i}(1-\rho_i),\ \rho_i = \frac{\lambda_i}{\mu_i}$$

Obviously, the joint distribution

$$p(n) = p_1(n_1)p_2(n_2)\cdots p_K(n_K) = \frac{1}{G}\rho_1^{n_1}\rho_2^{n_2}\cdots\rho_K^{n_K}$$

where G is the normalization factor. One easy verification is to check the detailed balance equations although we could use the independence to obtain it directly.

Truncation: a truncation of a system is a Markov chain having the same transition diagram with the only difference that some states are *eliminated*, while transitions between all other pairs of states with the same transition probabilities/rates.

Let S denote the set of states in the truncated queue system, a wild guess for the probability distribution would be

$$p(n) = \frac{1}{G}\rho_1^{n_1}\rho_2^{n_2}\cdots\rho_K^{n_K}$$

where G is the normalization factor over the state space of this truncated systems, i.e.,

$$G = \sum_{(n_1,\ldots,n_K)\in S} \rho_1^{n_1}\rho_2^{n_2}\cdots\rho_K^{n_K}.$$

Remark: This is very similar to the conditional probability argument.

Proof: Need to show that detailed balance equations

$$\lambda_j p(n_1, n_2, \ldots, n_{j-1}, n_j, n_{j+1}, \ldots, n_K) = \mu_j p(n_1, n_2, \ldots, n_{j-1}, n_j+1, n_{j+1}, \ldots, n_K).$$

In fact,

$$\begin{aligned} LHS &= \lambda_j \frac{\rho_1^{n_1}\cdots\rho_j^{n_j}\cdots\rho_K^{n_K}}{G} \\ &= \mu_j\rho_j \frac{\rho_1^{n_1}\cdots\rho_j^{n_j}\cdots\rho_K^{n_K}}{G} \\ &= \mu_j \frac{\rho_1^{n_1}\cdots\rho_j^{n_j+1}\cdots\rho_K^{n_K}}{G} \\ &= \mu_j p(n(j+)) = RHS \end{aligned}$$

Remarks:

(a). In the above proof, we must have $n(j+) \in S$, otherwise, it will lead us to the boundary condition.

(b). The difficulty: computation of G.

(c). Apply truncations to other cases such as $M/M/m$, $M/M/\infty$, $M/M/m/m$ as well.

Illustrative example (example 3.13)

Two session classes with preferential treatment for one class in a circuit switching system.

Consider a transmission line consisting of m independent circuits of equal capacity. Assuming that the two classes of sessions have arrival rates λ_1 and λ_2 and service rates μ_1 and μ_2, respectively. There is a limit $K < m$ on the # of circuits used by sessions of second type, thus there are at least (m-k) circuits to be used by the sessions of the first type. Interesting quantities in this system are the blocking probabilities:

- first type: blocked when all circuits are busy
- second type: blocked when there are K sessions of second type in the system

The Markov process $N(t) = (n_1(t), n_2(t))$ for this system is two-dimensional MC, where $n_i(t)$ is the number of customers of type i in the system at time t. It is easy to draw the state transition diagram (see Figure 3). The probability distribution is given by

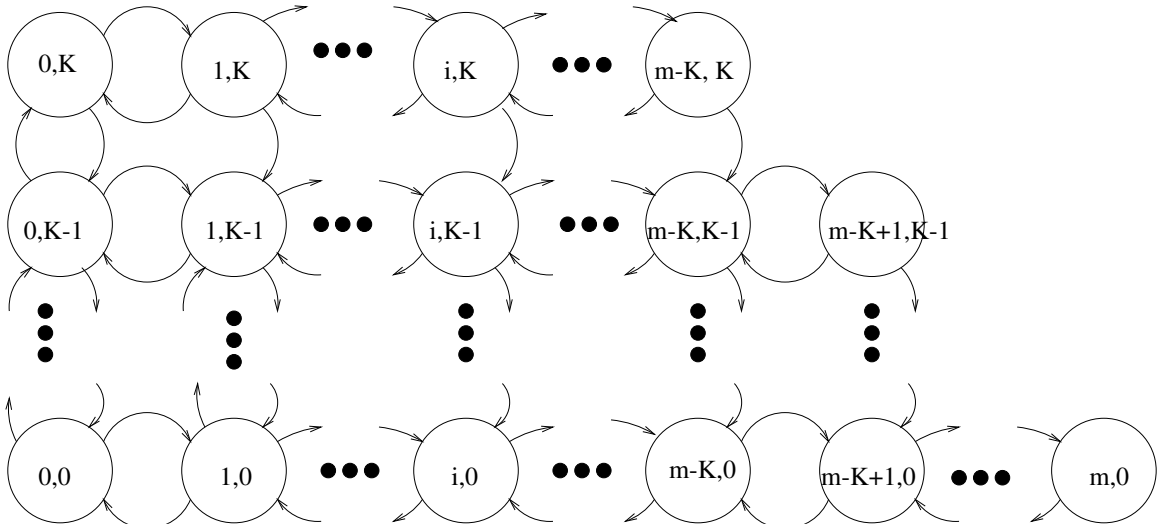


Figure 3: Multidimensional MC transition diagram

$$p(n_1, n_2) = \frac{\frac{\rho_1^{n_1}}{n_1!} \cdot \frac{\rho_2^{n_2}}{n_2!}}{G}$$

where

$$G = \sum_{(n_1,n_2)\in S} \frac{\rho_1^{n_1}}{n_1!} \cdot \frac{\rho_2^{n_2}}{n_2!}.$$

Verification: Check the detailed balance equations:

$$\begin{aligned} \lambda_1 p(n_1, n_2) &= (n_1+1)\mu_1 p(n_1+1, n_2) \\ \lambda_2 p(n_1, n_2) &= (n_2+1)\mu_2 p(n_1, n_2+1) \end{aligned}$$

In fact, we have

$$\begin{aligned} \lambda_1 p(n_1, n_2) &= \lambda_1 \frac{\frac{\rho_1^{n_1}}{n_1!} \cdot \frac{\rho_2^{n_2}}{n_2!}}{G} \\ &= \lambda_1 \cdot \frac{n_1+1}{\rho_1} \frac{\frac{\rho_1^{n_1+1}}{(n_1+1)!} \cdot \frac{\rho_2^{n_2}}{n_2!}}{G} = (n_1+1)\mu_1 p(n_1+1, n_2) \end{aligned}$$

The other equation can be proved similarly.

The normalization factor is given by

$$\begin{aligned} G &= \sum_{(n_1,n_2)\in S} \frac{\rho_1^{n_1}}{n_1!} \cdot \frac{\rho_2^{n_2}}{n_2!} \\ &= \sum_{n_1=0}^{m-K} \frac{\rho_1^{n_1}}{n_1!} \sum_{n_2=0}^{K} \frac{\rho_2^{n_2}}{n_2!} + \sum_{n_1=m-K+1}^{m} \sum_{n_2=0}^{m-n_1} \frac{\rho_1^{n_1}}{n_1!} \cdot \frac{\rho_2^{n_2}}{n_2!} \\ &= \left(\sum_{n_1=0}^{m-K} \frac{\rho_1^{n_1}}{n_1!} \right) \left(\sum_{n_2=0}^{K} \frac{\rho_2^{n_2}}{n_2!} \right) + \sum_{n_1=m-K+1}^{m} \sum_{n_2=0}^{m-n_1} \frac{\rho_1^{n_1}}{n_1!} \cdot \frac{\rho_2^{n_2}}{n_2!} \end{aligned}$$

Let P_{b1} and P_{b2} denote the blocking probabilities for the sessions of the first type and second type, respectively. Then we can obtain

$$\begin{aligned} P_{b1} &= \sum_{(n_1,n_2) \text{ is on diagonal}} p(n_1, n_2) = \sum_{n_1+n_2=m} p(n_1, n_2) \\ &= \frac{1}{G} \sum_{n_1=m-K} \frac{\rho_1^{n_1}}{n_1!} \cdot \frac{\rho_2^{m-n_1}}{(m-n_1)!} \\ P_{b1} &= \sum_{(n_1,n_2) \text{ is on outer boundary}} p(n_1, n_2) = \sum_{n_1+n_2=m \text{ or } n_2=K} p(n_1, n_2) \\ &= \frac{1}{G} \left[\left(\sum_{n_1=0}^{m-K} \frac{\rho_1^{n_1}}{n_1!} \right) \frac{\rho_2^K}{K!} + \sum_{n_1=m-K+1}^{m} \frac{\rho_1^{n_1}}{n_1!} \cdot \frac{\rho_2^{m-n_1}}{(m-n_1)!} \right] \end{aligned}$$

Remarks:

- There are mistakes in the textbook (on page 185).
- This result can be used in wireless networks.

8 Networks of Queues

Motivation: a few queues must be considered together because they are "related" or "interconnected": one's departures form another's arrivals. For example, a tandem of queues (in assembly lines), a cellular system of multiple cells, LANs on campus, $\cdots$

8.1 Open Queueing Networks (OQN)

Consider a network of K FCFS, single-server queues. Assume that

(1) external arrivals are Poisson with rate r_i;

(2) $\sum_i r_i > 0$;

(3) p_{ij} is the routing probability from node i to node j;

(4) a job leaves the network from node i with probability

$$1 - \sum_{j=1}^{K} p_{ij};$$

(5) each job/customer will eventually exit the system;

(6) service time in any queue is exponentially distributed.

Let λ_i denote the total arrival rate of jobs at queue i, then

$$\lambda_j = r_j + \sum_{i=1}^{K} \lambda_i p_{ij}, \quad j = 1, 2, \ldots, K \tag{12}$$

Let

$$\lambda = \begin{pmatrix} \lambda_1 \\ \lambda_2 \\ \vdots \\ \lambda_K \end{pmatrix}, \; r = \begin{pmatrix} r_1 \\ r_2 \\ \vdots \\ r_K \end{pmatrix}, \; P = \begin{pmatrix} p_{11} & p_{21} & \cdots & p_{K1} \\ p_{12} & p_{22} & \cdots & p_{K2} \\ \vdots & \vdots & \ddots & \vdots \\ p_{1K} & p_{2K} & \cdots & p_{KK} \end{pmatrix}, I = \begin{pmatrix} 1 & 0 & \cdots & 0 \\ 0 & 1 & \cdots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \cdots & 1 \end{pmatrix}$$

Then equations (12) becomes $(I - P)\lambda = r$. So that it has a unique solution iff $I - P$ is non-singular, iff 1 is not an eigenvalue of P, iff $\lim_{m\to\infty} P^m = 0$, which is implied by Assumption (5): for any i_1, there exists i with $1 - \sum_{j=1}^{K} p_{ij} > 0$ such that there exist $i_2, i_3, \ldots, i_k$ satisfying

$$p_{i_1 i_2} > 0, \; p_{i_2 i_3} > 0, \ldots, p_{i_k i} > 0.$$

Let $\rho_j = \lambda_j/\mu_j$, $j = 1, 2, \ldots, K$, assuming that $\rho_j < 1$.

Multi-dimensional Markov Chain (page 180)

State: $n = (n_1, n_2, \ldots, n_K)$
Problem: Find $p(n_1, n_2, \ldots, n_K)$
Notations:

$$\begin{aligned} n(j+) &= (n_1, \ldots, n_{j-1}, n_j + 1, n_{j+1}, \ldots, n_K) \\ n(j-) &= (n_1, \ldots, n_{j-1}, n_j - 1, n_{j+1}, \ldots, n_K) \\ n(i+, j-) &= (n_1, \ldots, n_i + 1, \ldots, n_j - 1, \ldots, n_K) \end{aligned}$$

Possible transitions (events)

- external arrival: $q_{nn(j+)} = r_j$
- external departure: $q_{nn(j-)} = \mu_j(1 - \sum_{i=1}^{K} p_{ji})$
- internal transition: $q_{nn(i+,j-)} = \mu_j p_{ji}$

Idea: Due to the randomization idea (routing probability formulation), each queue may be considered "independent"! If we regard all queues are "independent", then the probability at state $n = (n_1, n_2, \ldots, n_K)$ should be given by

$$p(n_1, n_2, \ldots, n_K) = p_1(n_1)p_2(n_2)\cdots p_K(n_K)$$

Jackson's Theorem: Under the assumptions made above, if $\rho_j < 1$, $j = 1, 2, \ldots, K$, we have for all $(n_1, n_2, \ldots, n_K) \geq 0$,

$$p(n) = p(n_1, n_2, \ldots, n_K) = p_1(n_1)p_2(n_2)\cdots p_K(n_K)$$

where

$$p_j(n_j) = \rho_j^{n_j}(1 - \rho_j),\ n_j \geq 0,\ j = 1, 2, \ldots, K$$

Proof: Idea: using an intelligent guess to get the stationary probability distribution as above, then verify the total balance equations via the time-reversibility theorem!

For any states n and n', let $q_{nn'}$ denote the corresponding rate. Define

$$q^*_{nn'} = \frac{p(n')q_{n'n}}{p(n)}$$

In order to show that $p(n)$ is the stationary distribution, we only need to show the total balance equation:

$$\sum_{n'} q_{nn'} = \sum_{n'} q^*_{nn'} \tag{13}$$

which is equivalent to

$$p(n) \sum_{n'} q_{nn'} = \sum_{n'} p(n') q_{n'n}$$

because $p(n) > 0$ and $\sum_n p(n) = 1$.

We first derive various transition rates, which are given in the following table:

forward chain	reversed chain
$q_{nn(j+)} = r_j$	$q^*_{nn(j+} = \frac{p(n(j+))q_{n(j+)n}}{p(n)} = \lambda_j(1 - \sum_i p_{ji})$
$q_{nn(j-)} = \mu_j(1 - \sum_i p_{ji})$	$q^*_{nn(j-)} = \frac{\mu_j r_j}{\lambda_j}$
$q_{nn(i+,j-)} = \mu_j p_{ji}$	$q^*_{nn(i+,j-)} = \frac{\mu_j \lambda_i p_{ij}}{\lambda_j}$
$q_{nn'} = 0$, other n'	$q^*_{nn'} = 0$, other n'

where

$$p(n(j+)) = \rho_j p(n),\ p(n(i+, j-)) = \rho_i p(n)/\rho_j.$$

Next, we verify balance equation (13):

$$\begin{aligned}
\sum_{n'} q_{nn'} &= \sum_{j=1}^{K} q_{nn(j+)} + \sum_{\{(j,i)|n_j>0\}} q_{nn(i+,j-)} + \sum_{\{j|n_j>0\}} q_{nn(j-)} \\
&= \sum_{j=1}^{K} r_j + \sum_{\{(j,i)|n_j>0\}} \mu_j p_{ji} + \sum_{\{j|n_j>0\}} \mu_j (1 - \sum_{i=1}^{K} p_{ji}) \\
&= \sum_{j=1}^{K} r_j + \sum_{\{j|n_j>0\}} \mu_j
\end{aligned} \tag{14}$$

$$\begin{aligned}
\sum_{n'} q^*_{nn'} &= \sum_{j=1}^{K} q^*_{nn(j+)} + \sum_{\{(j,i)|n_j>0\}} q^*_{nn(i+,j-)} + \sum_{\{j|n_j>0\}} q^*_{nn(j-)} \\
&= \sum_{j=1}^{K} \lambda_j (1 - \sum_{i=1}^{K} p_{ji}) + \sum_{\{(j,i)|n_j>0\}} \frac{\mu_j \lambda_i p_{ij}}{\lambda_j} + \sum_{\{j|n_j>0\}} \frac{\mu_j r_j}{\lambda_j} \\
&= \sum_{j=1}^{K} \lambda_j (1 - \sum_{i=1}^{K} p_{ji}) + \sum_{\{j|n_j>0\}} \frac{\mu_j [r_j + \sum_{i=1}^{K} \lambda_i p_{ij}]}{\lambda_j} \\
&= \sum_{j=1}^{K} \lambda_j (1 - \sum_{i=1}^{K} p_{ji}) + \sum_{\{j|n_j>0\}} \frac{\mu_j \lambda_j}{\lambda_j}
\end{aligned}$$

$$= \sum_{j=1}^{K} \left[\lambda_j - \sum_{i=1}^{K} \lambda_j p_{ji} \right] + \sum_{\{j|n_j>0\}} \mu_j$$
$$= \sum_{j=1}^{K} r_j + \sum_{\{j|n_j>0\}} \mu_j$$

which is equal to the right hand side of equation (14), hence the total balance equation holds, therefore $p(n)$ is the stationary distribution. This proves Jackson's Theorem.

Remarks:

(1). $q^*_{nn'}$ is the transition rates for the reversed process, which corresponds to a network of queues where traffic arrives at queue i from outside the network according to a Poisson process with rates $\lambda_i(1 - \sum_j p_{ij})$, with the routing probabilities

$$\frac{\lambda_j p_{ji}}{r_i + \sum_k \lambda_k p_{ki}} = \frac{\lambda_j p_{ji}}{\lambda_i}$$

which is also the probability that an arriving customer at queue i just departed from queue j in the forward process.

(2). The interarrival time in the reversed system are independent and exponentially distributed, i.e., the departure process in the forward process is Poisson!

(3). In Jackson networks, the number of jobs in the system's queues are distributed as if each queue is $M/M/1$ and is independent of the other queues. However, the total arrival process at each queue needs not be Poissonian!

Example 3.19. Consider system with feedback loop for I/O as shown in Figure 4). Assume all

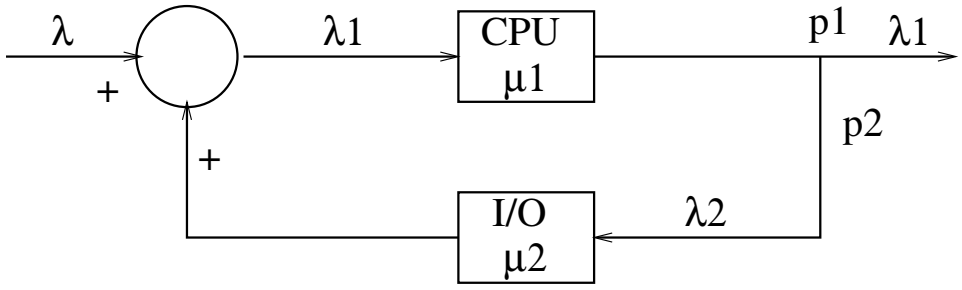


Figure 4: Figure for Example 3.19

"Jackson's assumptions" hold, find the probability distribution $p(n_1, n_2)$.

Solution: Write down the equations:

$$\begin{aligned} \lambda_1 &= \lambda + \lambda_2 \\ \lambda_2 &= p_2\lambda_1 \end{aligned}$$

with $p_1 + p_2 = 1$. Solving the equations we obtain the solution

$$\begin{aligned} \lambda_1 &= \frac{\lambda}{p_1} > 0 \\ \lambda_2 &= \frac{p_2}{p_1} > 0 \end{aligned}$$

Define

$$\begin{aligned} \rho_1 &= \frac{\lambda_1}{\mu_1} = \frac{1}{p_1}\left(\frac{\lambda}{\mu_1}\right) \\ \rho_2 &= \frac{\lambda_2}{\mu_2} = \frac{p_2}{p_1}\left(\frac{\lambda}{\mu_2}\right) \end{aligned}$$

From Jackson's Theorem, we obtain

$$\begin{aligned} p_1(n_1) &= \rho_1^{n_1}(1-\rho_1) \\ p_2(n_2) &= \rho_2^{n_2}(1-\rho_2) \end{aligned}$$

8.2 Extensions of Jackson's Theorem

There are many variations of Jackson networks under which Jackson's Theorem is still valid. We here only discuss a few.

8.2.1 State-dependent Service Rates

Multiservices case: when the service rate depends on the state of the queue, say, the service time is exponentially distributed with rate $1/\mu_j(m)$ when $n_j = m$. Let

$$\rho_j(m) = \frac{\lambda_j}{\mu_j(m)}$$

where

$$\lambda_j = r_j + \sum_i \lambda_i p_{ij}.$$

Define

$$\hat{p_j}(n_j) = \begin{cases} 1 & \text{if } n_j = 0 \\ \rho_j(1)\rho_j(2)\cdots\rho_j(n_j) & \text{if } n_j > 0 \end{cases}$$

Jackson's Theorem: Under Jackson's assumptions, for any $n = (n_1, n_2, \ldots, n_K)$, we have

$$p(n) = \frac{1}{G}\hat{p_1}(n_1)\hat{p_2}(n_2)\cdots\hat{p_K}(n_K)$$

assuming that $0 < G < \infty$, where

$$G = \sum_{n_1}^{\infty}\sum_{n_2=0}^{\infty}\cdots\sum_{n_K=0}^{\infty}\hat{p_1}(n_1)\hat{p_2}(n_2)\cdots\hat{p_K}(n_K),$$

the so-called normalization factor.

8.2.2 Multiple Classes of Customers

Let $c = 1, 2, \ldots, C$ denote the classes of jobs/customers, let $r_j(c)$ denote the rate of the external Poisson arrivals of class c jobs at node j. Let $p_{ij}(c)$ denote the routing probability of class c job routing from node i to node j. Assume that

$$\lambda_j(c) = r_j(c) + \sum_{i=1}^{K}\lambda_i(c)p_{ij}(c),\ j = 1, 2, \ldots, K, c = 1, 2, \ldots, C$$

has a unique solution $\lambda_j(c)$ such that $\lambda_j(c) > 0$. Assume also that the service times at queue j are exponentially distributed with rate $\mu_j(m)$ for *all customer classes*. In this system, the state is a composition of all queue states:

$$z = (z_1, z_2, \ldots, z_K)$$

where

$$z_j = (c_1, c_2, \ldots, c_{n_j})$$

with

$$\begin{aligned} n_j &= \text{\# of customers in the } j\text{th queue} \\ c_j &= \text{the class number of the customers in the } i\text{th queue position} \end{aligned}$$

Define

$$\begin{aligned} \hat{\rho_j}(c, m) &= \frac{\lambda_j(c)}{\mu_j(m)},\ j = 1, 2, \ldots, K, c = 1, 2, \ldots, C \\ \hat{\rho_j}(z_j) &= \begin{cases} 1 & \text{if } n_j = 0 \\ \hat{\rho_j}(c_1, 1)\hat{\rho_j}(c_2, 2)\cdots\hat{\rho_j}(c_{n_j}, n_j) & \text{if } n_j > 0 \end{cases} \\ G &= \sum_{(z_1, z_2, \ldots, z_K)}\prod_{j=1}^{K}\hat{p_j}(z_j)\ \text{ (normalization factor)} \end{aligned}$$

Jackson's Theorem: Assuming that $0 < G < \infty$, then the steady-state probability $\hat{p}(z)$ of the state $z = (z_1, z_2, \ldots, z_K)$ is given by the following product form:

$$\hat{p}(z) = \frac{1}{G}\hat{p_1}(z_1)\hat{p_2}(z_2)\cdots\hat{p_K}(z_K).$$

The steady-state probability $p(n) = p(n_1, n_2, \ldots, n_K)$ of having a total of n_j customers at queue $j = 1, 2, \ldots, K$ is given by

$$p(n) = \sum_{z \in Z(n)} \hat{p}(z)$$

where $Z(n)$ is the set of states for which there exists a total of n_j customers in queue j.

Example: When $C = 1$, we obtain the single class OQN.

8.3 The Kleinrock Independence Approximation: Virtual Circuit Networks

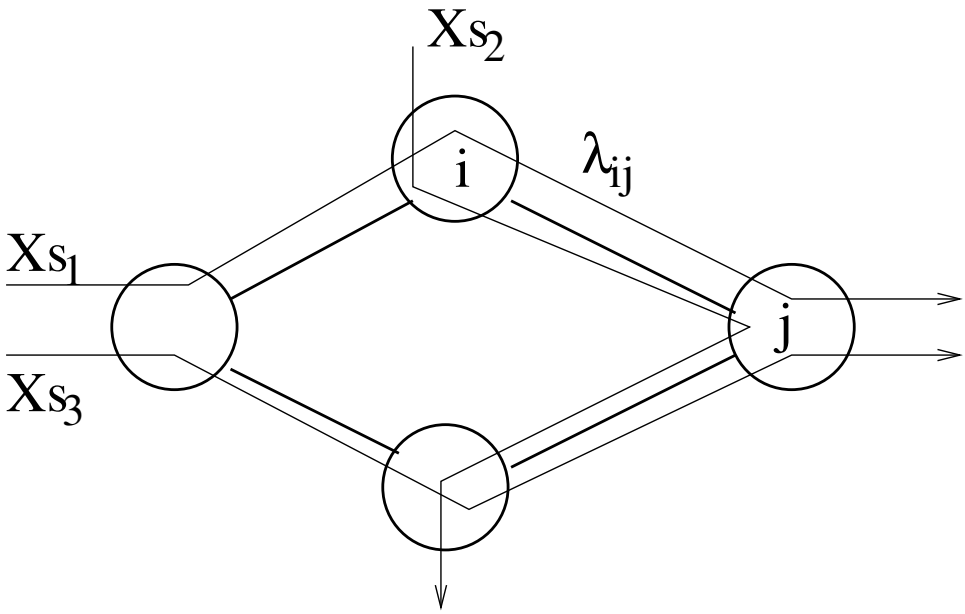


Figure 5: Kleinrock Independence

Assume that there exist several packet streams in the open network, each stream following a unique path that consists of multiple links through the network (see Figure 5). Let X_s be the arrival rate of packet stream s. The total arrival rate at link (i, j) is

$$\lambda_{ij} = \sum_{\text{all packet streams } s \text{ crossing } (i,j)} X_s$$

If packet flow bifurcation is allowed, i.e., $f_{ij}(s)$: the fraction of the flow s traversing (i,j) link, so

$$\lambda_{ij} = \sum_{s \to (i,j)} f_{ij}(s) X_s.$$

Idea: If over the link, there are substantial amount of external Poisson traffic injected, the overall traffic over the link can be modeled by the Poisson process, leading to the following approximation: the Poisson dominant traffic will represent the overall traffic (see Figure 6). If the service at each

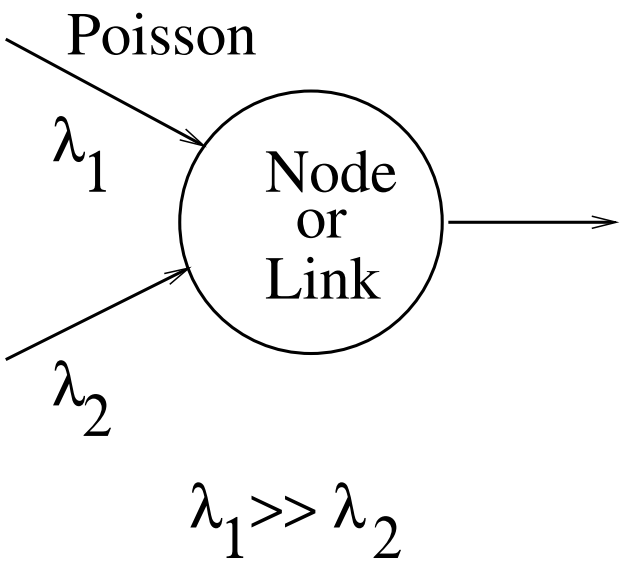


Figure 6: Approximation illustration

link is exponentially distributed, then each link can be modeled as $M/M/1$. This is known as the *Kleinrock Independence Approximation* (in terms of interarrival times).

Under this condition, let $1/\mu_{ij}$ denote the average transmission time over link (i,j), N_{ij} denotes the average number over link (i,j) and N the average number in the network, then

$$\begin{aligned} N_{ij} &= \frac{\lambda_{ij}}{\mu_{ij} - \lambda_{ij}} \\ N &= \sum_{(i,j)} N_{ij} = \sum_{(i,j)} \frac{\lambda_{ij}}{\mu_{ij} - \lambda_{ij}} \end{aligned}$$

From Little's law, the average system (network) time or delay is

$$T = \frac{N}{\gamma} = \frac{\sum_{(i,j)} \frac{\lambda_{ij}}{\mu_{ij} - \lambda_{ij}}}{\sum_s X_s}.$$

8.4 Network Delay in Queueing Networks

In Open Queueing Networks, we observe that the Kleinrock Independence Approximation is also valid: each node can be modeled as $M/M/1$, thus,

$$\begin{aligned} N_i &= \frac{\lambda_i}{\mu_i - \lambda_i} \\ N &= \sum_i N_i = \sum_i \frac{\lambda_i}{\mu_i - \lambda_i} \\ T &= \frac{N}{\lambda} = \frac{\sum_i \frac{\lambda_i}{\mu_i - \lambda_i}}{\sum_{i=1}^{K} r_i} \end{aligned}$$

Another approach:

$$\begin{aligned} \lambda T &= NE[n] = \sum_{(n_1, n_2, \ldots, n_K)} (n_1 + n_2 + \cdots + n_K) p(n_1, n_2, \ldots, n_K) \\ &= \sum_i \left[\sum_{n_i} n_i \sum_{n_1, \ldots, n_{i-1}, n_{i+1}, \ldots, n_K)} p(n_1, n_2, \ldots, n_K) \right] \\ &= \sum_i \left[\sum_{n_i} n_i p_i(n_i) \right] = \sum_i E[n_i] = \sum_i N_i \end{aligned}$$

End-to-end Delay

$$E[T_{i_1 i_2 \cdots i_r}] = \sum_m T_{i_m} = \sum_m \frac{1}{\mu_{i_m} - \lambda_{i_m}}.$$

8.5 Closed Queueing Networks (CQN)

The total number of jobs remains constant!

- M— # of customers/jobs in the system
- p_{ij}— routing probability from queue i to queue j and $\sum_{j=1}^{K} p_{ij} = 1$
- $\mu_j(m)$— service rate at jth queue when there exists m jobs
- λ_j— total arrival rate to queue i

Thus, we must have

$$\begin{aligned} \lambda_j &= \sum_{i=1}^{K} \lambda_i p_{ij},\ j = 1, 2, \ldots K \\ \lambda &= P\lambda,\ (I - P)\lambda = 0 \end{aligned} \tag{15}$$

where $P = (p_{ij})$ is the routing matrix and I is the identity matrix. Assume that P is irreducible (i.e., if we regard p_{ij} as the transition probabilities for a finite-state Markov chain, the Markov chain is irreducible). Then, all solutions $\lambda = (\lambda_1, \lambda_2, \ldots, \lambda_K)^T$ of equation (15) will be in the form of

$$\lambda_j = \alpha \bar{\lambda}_j,\ j = 1, 2, \ldots, K$$

where α is a scalar and $\bar{\lambda}_j$ ($j = 1, 2, \ldots, K$) is a particular solution with $\bar{\lambda}_j > 0$ for all j. Let

$$\begin{aligned} \rho_j(m) &= \frac{\bar{\lambda}_j}{\mu_j(m)},\ j = 1, 2, \ldots, K, m = 1, 2, \ldots \\ \hat{p}_j(n_j) &= \begin{cases} 1 & \text{if } n_j = 0 \\ \rho_j(1)\rho_j(2)\cdots\rho_j(n_j) & \text{if } n_j > 0 \end{cases} \\ G &= G(M) = \sum_{n_1+n_2+\cdots+n_K=M} \hat{p}_1(n_1)\hat{p}_2(n_2)\cdots\hat{p}_K(n_K) \end{aligned}$$

Jackson's Theorem for CQN: Under the proceeding assumptions, for any $n = (n_1, n_2, \ldots, n_K)$ satisfying $n_1 + n_2 + \cdots + n_K = M$, we have

$$p(n) = \frac{1}{G}\hat{p}_1(n_1)\hat{p}_2(n_2)\cdots\hat{p}_K(n_K).$$

Proof: Similar to the proof for OPN, for any n, n', define

$$q^*_{nn'} = \frac{p(n')q_{n'n}}{p(n)}$$

we need to prove the total balance equations. We use the same notation, we can easily verify the following:

$$\begin{aligned} q_{nn(i+,j-)} &= \mu_j(n_j)p_{ji} \\ q^*_{nn(i+,j-)} &= \frac{p(n(i+,j-))q_{n(i+,j-)n}}{p(n)} \\ &= \frac{\rho_i(n_i+1)}{\rho_j(n_j)}\mu_i(n_i+1)p_{ij} = \frac{\mu(n_j)\bar{\lambda}_i}{\bar{\lambda}_j}p_{ij} \end{aligned}$$

Now, we verify the total balance equation

$$\sum_m q_{nm} = \sum_m q^*_{nm}.$$

In fact, we can easily obtain

$$\begin{aligned}\sum_m q_{nm} &= \sum_{(j,i)|n_j>0} q_{nn(i+,j-)} = \sum_{(j,i)|n_j>0} \mu_j(n_j) p_{ji} \\ &= \sum_{i=1}^{K} \sum_{n_j>0} \mu_j(n_j) p_{ji} = \sum_{n_j>0} \mu_j(n_j) \left(\sum_{i=1}^{K} p_{ji} \right) \\ &= \sum_{n_j>0} \mu_j(n_j)\end{aligned}$$

and

$$\begin{aligned}\sum_m q^*_{nm} &= \sum_{(j,i)|n_j>0} q^*_{nn(i+,j-)} = \sum_{(j,i)|n_j>0} \frac{\mu_j(n_j)\bar{\lambda}_i p_{ij}}{\bar{\lambda}_j} \\ &= \sum_{n_j>0} \frac{\mu_j(n_j)}{\bar{\lambda}_j} \left(\sum_{i=1}^{K} \bar{\lambda}_i p_{ij} \right) = \sum_{n_j>0} \mu_j(n_j)\end{aligned}$$

This completes the proof.

Remark: Obviously, the probability distribution $p(n)$ does not depend on the choice of the solution of (15) because all solutions are just a constant multiple of the particular solution, the constant has been cancelled out in the probability distribution.

Example: Consider the CQN shown in Figure 7. First, we solve the rate equations

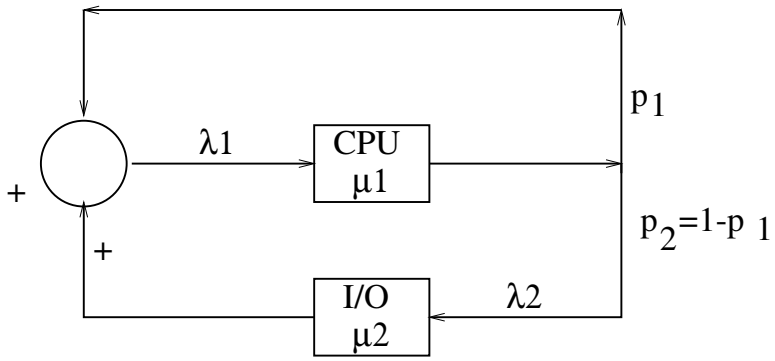


Figure 7: CQN example

$$\begin{aligned}\lambda_1 &= p_1\lambda_1 + \lambda_2 \\ \lambda_2 &= p_2\lambda_1\end{aligned}$$

Choose $\bar{\lambda}_1 = \mu_1$ and $\bar{\lambda}_2 = p_2\mu_1$, we obtain

$$\rho_1 = 1,\ \rho_2 = \frac{p_2\mu_1}{\mu_2}.$$

Then, the steady-state probability is given by

$$p(M-n,n) = \frac{1}{G}\rho_2^n,\ n = 0,1,2,\ldots,M,$$

where

$$G(M) = G = \sum_n \hat{p}_1(n)\hat{p}_2(n) = \sum_n \rho_2^n = \frac{1-\rho_2^{M+1}}{1-\rho_2}.$$

The utilization factor for the CPU is given by

$$U(M) = 1 - p(0,M) = 1 - \frac{\rho_2^M}{G(M)} = \frac{G(M)-\rho_2^M}{G(M)} = \frac{G(M-1)}{G(M)}.$$

The arrival rate at CPU is (from the Little's law)

$$\lambda_1(M) = U(M)\mu_1.$$

8.5.1 Computation of Normalization Factor G

Although the result for CQN is beautiful in nature, the difficulty in using it is how to compute the normalization factor $G = G(M)$. When M is large, we may face the curse-of-dimensionality problem! In this subsection, we present a recursive algorithm for a special case: when the service does not depend on the number of customers, i.e., $\mu_i(m) = \mu_i$. Details can be found in Schwartz's book (page 225).

Let

$$G(n,k) = \sum_{n_1+n_2+\cdots+n_k=n} \prod_{i=1}^{k} \rho_i^{n_i}$$

where $\rho_i = \bar{\lambda}_j/\mu_j(m) = \bar{\lambda}_j/\mu_j$. We claim that all desired statistics of the CQN can be obtained from the normalization factor $G(M,K)$. The normalization factor can be obtained from the following recursive formula

$$G(n,k) = G(n,k-1) + \rho_k G(n-1,k),\ \rho_k = \lambda_k/\mu_k \tag{16}$$

with the initial starting conditions given by

$$\begin{aligned} G(n,1) &= \rho_1^n,\ n = 1,2,\ldots,M \\ G(0,k) &= 1,\ k = 1,2,\ldots,K \end{aligned}$$

Proof: We can divide the summation into two cases: when $n_k = 0$ and $n_k > 0$, so we have

$$G(n,k) = \sum_{n_1+\cdots+n_k=n} \prod_{i=1}^{k} \rho_i^{n_i}$$

$$
\begin{aligned}
&= \sum_{n_1+\cdots+n_{k-1}=n} \prod_{i=1}^{k-1} \rho_i^{n_i} + \sum_{n_1+\cdots+n_k=n, n_k>0} \left(\prod_{i=1}^{k-1} \rho_i^{n_i}\right) \rho_k^{n_k} \\
&= G(n,k-1) + \rho_k \sum_{n_1+\cdots+(n_k-1)=n-1, n_k>0} \left(\prod_{i=1}^{k-1} \rho_i^{n_i}\right) \rho_k^{n_k-1} \\
&= G(n,k-1) + \rho_k \sum_{n_1+\cdots+n_k=n-1} \left(\prod_{i=1}^{k-1} \rho_i^{n_i}\right) \rho_k^{n_k} \\
&= G(n,k-1) + \rho_k G(n-1,k)
\end{aligned}
$$

This completes the proof.

Next, we show that many statistics can be obtained from the normalization factor function.

(i). $\Pr(n_i \geq k) = \rho_i^k G(M-k,K)/G(M,K)$.

Proof:

$$
\begin{aligned}
\Pr(n_i \geq k) &= \sum_{n_1+\cdots+n_K=M,\ n_i\geq k} p(n) \\
&= \sum_{n_1+\cdots+n_K=M,\ n_i\geq k} \frac{\prod_{j=1}^K \rho_j^{n_j}}{G(M,K)} \\
&= \frac{\rho_i^k \left(\sum_{n_1+\cdots+(n_i-k)+\cdots+n_K=M-k} \left(\prod_{j\neq i} \rho_j^{n_j}\right) \rho_i^{n_i-k}\right)}{G(M,K)} \\
&= \frac{\rho_i^k G(M-k,K)}{G(M,K)}
\end{aligned}
$$

(ii). Marginal distribution:

$$
\Pr(n_i = k) = \Pr(n_i \geq k) - \Pr(n_i \geq k+1) = \frac{\rho_i^k [G(M-k,K) - \rho_i G(M-k-1,K)]}{G(M,K)}.
$$

(iii). Average number of jobs at a node:

$$
\begin{aligned}
E(n_i) &= \sum_{k=0}^M k \Pr(n_i = k) = \sum_{k=0}^M k[\Pr(n_i \geq k) - \Pr(n_i \geq k+1)] \\
&= \sum_{k=0}^M \Pr(n_i \geq k) = \sum_{k=1}^M \rho_i^k \frac{G(M-k,K)}{G(M,K)}
\end{aligned}
$$

(iv). Average total number

$$
\begin{aligned}
N &= \sum_{i=1}^K E(n_i) = \sum_{i=1}^K \sum_{k=1}^M \rho_i^k \frac{G(M-k,K)}{G(M,K)} \\
&= \sum_{r=1}^n \left(\sum_{i=1}^M \rho_i^r\right) \frac{G(M-r,K)}{G(M,K)} = M
\end{aligned}
$$

Remark: In any case, recursive procedure can be developed for almost all statistics.

8.6 Important Generalizations

Intensive research has been carried out for finding conditions under which the product form exists. There are a few special classes:

- Phase-type distributed service time: (1) processor sharing ($\mu \rightarrow \mu/n$); (2) last come first serve (LCFS)
- BCMP: most general set of conditions under which the product form exists.

References

- F.P. Kelly, *Reversibility and Stochastic Networks*, John Wiley and Sons, 1979.
- F. Baskett, M. Chandy, R. Muntz and J. Palacios, "Open, closed and mixed networks of queues with different classes of customers," *Journal of the ACM*, **22**, 248-260, 1975.

The details will be discussed in the future.

9 BCMP Networks

The pursuit of product form for queueing networks leads to BCMP networks. People are searching for conditions/assumptions under which the product form will be valid, service time distribution and service disciplines seem to be the key factors!

9.1 Service Time Distributions

Exponential distribution has been used in many applications due to its simplicity, however, many time variables we come across today show non-exponential characteristics: service is fat-tailed (web traffic), packet arrivals show the break-down of Poisson traffic. More general models are needed.

Search for the model:

- General enough to approximate the field data
- Simple enough to give the attractability for analytical solutions: the preservation of Markov property

9.1.1 Staging methods (Cox, 1959)

Serial staging

Multiple exponential service facilities are connected in Tandem. Mathematically, we have

$$\xi = t_1 + t_2 + \cdots + t_m$$

where t_i is exponentially distributed with parameter μ_i. We have ($f^*(s)$ indicates the Laplace transform of $f(t)$)

$$\begin{aligned} E[\xi] &= \sum_{i=1}^{m} E[t_i] = \sum_{i=1}^{m} \frac{1}{\mu_i} \\ f_\xi^*(s) &= \left(\frac{\mu_1}{s+\mu_1}\right)\left(\frac{\mu_2}{s+\mu_2}\right)\cdots\left(\frac{\mu_m}{s+\mu_m}\right) \end{aligned}$$

Special cases:

- Erlang distribution: $\mu_1 = \mu_2 = \cdots = \mu_m = \mu$;

- Exponential distribution: $m = 1$.

Parallel staging

Traffic is splitting into multiple streams, each of which is an exponential service facility, the branching probabilities are α_i, $i = 1, 2, \ldots, m$ with $\sum_{i=1}^{m} \alpha_i = 1$. This is equivalent to the random variable ξ defined as $\xi = t_i$ with probability α_i, where t_i is exponentially distributed with parameter α_i. The Laplace transform is

$$f_\xi^*(s) = \sum_{i=1}^{m} \alpha_i \frac{\mu_i}{s + \mu_i}$$

which is also called *hyper-exponential* distribution.

Serial-Parallel Staging

The first stage is to branch with probabilities α_i, $i = 1, 2, \ldots, m$, each branch forms a serial staging, which gives the most general distribution. Assuming that the ith branch contains exponential service facilities with parameters: μ_{ij}, $j = 1, 2, \ldots, m_i$, then the Laplace transform is given by

$$f_\xi^*(s) = \sum_{i=1}^{m} \alpha_i \prod_{j=1}^{m_i} \left(\frac{\mu_{ij}}{s + m_{ij}} \right).$$

Special cases:

- Hyper-exponential distribution: $m_1 = m_2 = \cdots = m_m = 1$;
- Mixed-Erlang (hyper-Erlang): $\mu_{i1} = \mu_{i2} = \cdots = \mu_{im_i} = \mu_i$

  $$f_\xi^*(s) = \sum_{i=1}^{m} \alpha_i \left(\frac{\mu_i}{s + \mu_i} \right)^{m_i}$$

  or

  $$f_\xi^*(s) = \sum_{i=1}^{m} \alpha_i \left(\frac{m_i \mu_i}{s + m_i \mu_i} \right)^{m_i}$$

  where the latter can make the choice of μ_i independent of m_i.
- SOHYP (Sum of Hyper-exponential) distribution (Rappaport)

  $$f_\xi^*(s) = \prod_{i=1}^{m} \left(\sum_{j=1}^{m_i} \alpha_{ij} \frac{\mu_{ij}}{s + \mu_{ij}} \right).$$

Coxian model

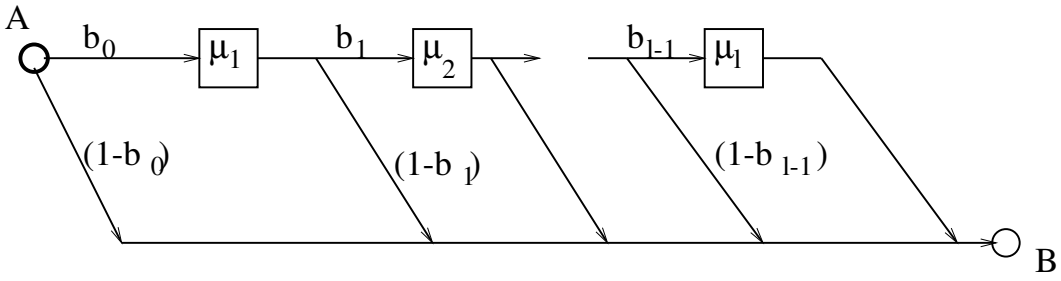


Figure 8: Coxian model

The service point consists of a set of exponential servers, arranged as shown in Figure 8. A new customer can enter into service only when the preceeding customer has left the staged network. Let the total number of servers in the Coxian distribution is l, the service time of the mth stage of Coxian model, having a exponential server with parameter μ_m. Define $A_m = b_0 b_1 \cdots b_{m-1}$, which can be interpreted as the probability that a customer reaches the mth stage, then the Laplace transform of the Coxian distribution is given by

$$f_\xi^*(s) = b_0 + \sum_{m=1}^{l} A_m(1 - b_m) \prod_{j=1}^{m} \left(\frac{\mu_j}{s + \mu_j} \right).$$

Phase-type distribution

This is a model general distribution model, which will play a very important role in matrix geometric approach. We will postpone such discussion later.

Remarks:

- All distribution models obtained from staging method preserve the Markovian property (in a multidimensional state space);
- Cox distributions are identical to distributions which have a rational Laplace transform;
- Erlang, hyper-exponential and series-parallel combinations are all Coxian;
- A sum of Cox distributions is also a Cox distribution.

9.2 Sevice disciplines

Service stations in a BCMP (basket, Chandy, Muntz, Palacios) network can obey any of the following possibilities:

Type 1: The service discipline is first in, first out (FIFO), the station has a single server and the service time is exponentially distributed with the same mean for all classes of customers. If station i has a such server, we denote the rate of service by $\mu_i(k_i)$ if there are k_i customers in the station (including the one which is in service).

Type 2: The discipline is that of time division ("processor sharing", PS), that is, a customer at the station receives $1/k$ seconds of service per second if there are k customers at the station. All customers receive a small portion of their respective servuce time in turn. These quanta of service received on each visit to the server tend to zero. The service time distribution can be a distinct Cox distribution for each class of customers.

Type 3: The number of servers at the station is sufficient for there to be always at least one free. This leads to the fact that a new customer entering the station starts his service immediately. The service time distributions can be distinct Cox distributions for each class of customers.

Type 4: The service discipline is "last in, first out" (LIFO) with an absolute priority for the newly arriving customer. There is a single server, that is a new arrival at the station interrupts the customer's service in order to start his own. The displaced customer is returned to the head of the queue and he rejoins his service where it left off, when the customer who caused the interruption finishes its service. The service time distribution can be Cox, which may be different for each class of customers.

9.3 BCMP Theorem

Let $i = 1, 2, \ldots, I$ be the indices of a partition of the classes of customers, $c = 1, 2, \ldots, R$. Let K_i denote the number of customers in the element of the partition having index i. Assuming that the arrival process of new customers in the network is Poisson with the following two possibilities:

- $\lambda(K)$ is the rate of arrivals from the exterior when there are K customers in the network;
- $\lambda_i(K_i)$ is the rate of arrivals from the exterior of customers in partition i of which there are K_i custoers in the network.

Let e_{ir} denote the relative frequency of the number of visits to station i by a customer of class r, then we have

$$e_{ir} = \sum_{r'=1}^{R} \left(\sum_{j=1}^{N} e_{jr'} p_{jr',ir} + p_{0r',ir} \right).$$

Let $\hat{k} = (\hat{k}_1, \ldots, \hat{k}_N)$ be the state vector of the network which depends on the type of service discipline:

Type 1: $\hat{k}_i = (k_{i1}, \ldots, k_{ik_i})$ where k_{ij} is the class of the jth customer waiting at station i in FCFS order.

Type 2 or 3:

$$\hat{k}_i = ((k_{i1}, s_{i1}), \ldots, (k_{ik_i}, s_{ik_i}))$$

where k_{ij} is the class of the jth customer waiting in the order of arrival and s_{ij} is the stage of the Cox model.

Type 4:

$$\hat{k}_i = ((k_{i1}, s_{i1}), \ldots, (k_{ik_i}, s_{ik_i}))$$

where k_{ij} and s_{ij} are identical to those defined above, the order of the k_{ij} being defined by LIFO discipline.

Define

$$f_i(\hat{k}_i) = \begin{cases} \prod_{j=1}^{k_i} \frac{e_{ik_{ij}}}{\mu_i(j)} & \text{if } i \text{ is of type 1} \\ k_i! \prod_{r=1}^{R} \prod_{m=1}^{l_r} \frac{(e_{ir}A_{irm}/\mu_{irm})^{k_{irm}}}{k_{irm}!} & \text{if } i \text{ is of type 2} \\ \prod_{r=1}^{R} \prod_{m=1}^{l_{ir}} \frac{(e_{ir}A_{irm}/\mu_{irm})^{k_{irm}}}{k_{irm}!} & \text{if } i \text{ is of type 3} \\ \prod_{j=1}^{k_i} \frac{e_{ik_{ij}}A_{ik_{ij}}}{\mu_{i(k_{ij},s_{ij})}} & \text{if } i \text{ is of type 4} \end{cases}$$

where parameters such as A_{irm}, $A_{ik_{ij}}$, l_{ir} and μ_{irm} are parameters in Cox distributions. Define

$$d(K) = \begin{cases} \prod_{m=0}^{K-1} \lambda(m) & \text{if the network is open} \\ 1 & \text{if the network is closed} \end{cases}$$

In the case of distinct arrival process for the sub-chains $1, 2, \ldots, n$, we define

$$d(K) = \begin{cases} \prod_{j=1}^{n} \prod_{i=0}^{K_i-1} \lambda(i) & \text{if the network is open} \\ 1 & \text{if the network is closed} \end{cases}$$

Define

$$\gamma = \sum_{\hat{k}} d(K) \prod_{i=1}^{N} f_i(\hat{k}_i)$$

BCMP Theorem: If $\gamma < +\infty$, a steady-state probability distribution exists and is given by

$$p(\hat{k}) = \gamma^{-1} d(K) \prod_{i=1}^{N} f_i(\hat{k}_i).$$

Remarks:

- The probability distribution involves only the first moments of the service distributions.
- The joint probability distribution is the product form of the marginal probabilities, which allows the network to be studied station by station.
- The principal difficulty encountered is calculation of the normalization constant γ.